School Quality and Massachusetts Enrollment Shifts in the Context of Tax Limitations

Katharine L. Bradbury, Karl E. Case, and Christopher J. Mayer

The authors are Vice President and Economist, Federal Reserve Bank of Boston; Professor of Economics, Wellesley College, and Visiting Scholar, Federal Reserve Bank of Boston; and Assistant Professor, Columbia Business School, respectively. *The authors owe thanks to Peter Fortune, Yolanda Kodrzycki, Eric Rosengren, and Robert Tannenwald for helpful comments, and to Jenny Liu for excellent research assistance.*

Like most states, Massachusetts underwent a large shift in public school enrollments between the 1980s and 1990s, requiring a number of sizable fiscal and educational adjustments by individual school districts. Between school years 1980 and 1989, the number of students in kindergarten through grade 12 fell 21 percent, from 1.04 million to 825,000. As children of baby boomers reached school age, the picture changed and enrollments grew more than 90,000 over the next seven years. These aggregate trends gloss over even more marked shifts at the local level.

Consider the communities of Brookline and Arlington, whose public school enrollments in school year 1980 were 6,246 and 6,245, respectively. Both are suburban communities located close to downtown Boston with little buildable land. The quality of Arlington's schools is considered slightly above average for the state, while Brookline's schools are perennially ranked among the top districts in the Commonwealth. By the mid 1990s the enrollment patterns for the two districts could not have looked more different. Arlington was closing schools. Despite the pickup in aggregate statewide enrollments, its 1996 enrollment was 4,059, a drop of more than one-third from 1980. Meanwhile Brookline experienced a much smaller decline in its number of students in the 1980s and faced an influx of students in the 1990s. By 1996, its enrollment was 6,039, barely 3 percent below its level in 1980.

This disparate pattern of enrollment shifts was not unique to these two communities. During the same 16-year period, almost one-quarter of Massachusetts communities lost more than 20 percent of students from their 1980 levels; at the other extreme, one-quarter gained more than 12 percent. These shifts in enrollment posed a significant fiscal challenge for communities struggling to provide facilities and teachers for the widely varying numbers of students. After all, educational expenditures represent almost one-half of the local budget for a typical community in Massachusetts.

The fact that households move is not surprising. Economists since Tiebout (1956) have recognized that households sort themselves based on their ability to pay and their preferences regarding both local public services and local housing characteristics, and these preferences can change over time as families begin having children or households decide to retire. Location models would predict a "flight to quality," for example, as households with children who reach school age choose to move to communities with higher-quality schools.

Shifts in school enrollment can pose a significant fiscal challenge: Educational expenditures represent almost one-half of the local budget for a typical community in Massachusetts.

As documented below, this pattern of sorting greatly increased between the 1980s and the 1990s in Massachusetts, with a much higher percentage of households with children moving in recent years. Demographics may explain part of this pattern. In the 1980s aggregate school enrollments were declining as the tail end of the baby boom was exiting the public schools and many older baby boomers had delayed childbearing. In the 1990s school enrollments were again rising. In addition, baby-boomers who were having children in the 1990s had additional income to spend on housing based on gains made in the housing market from the 1980s.

As households change their desires based on life-cycle considerations, economic models also predict that communities would adjust the amount of public services (such as police, fire, and schools) in response to the changing desires of households. Thus, one might have expected cities and towns to respond to the demographically driven increase in demand for good schools in the 1990s by raising educational expenditures. However, a statewide property tax limitation measure, Proposition 2½, raised strong barriers to providing desired services in some communities.

This article investigates the degree to which the constraints of Proposition 2½, and other factors such as demographic and economic shifts and differences in school quality, affected the adjustments that both local governments and households in the Commonwealth made to a demographically driven turnaround in enrollment growth. The research accomplishes this task by comparing changes in enrollments in the first half of the 1980s to those in the first half of the 1990s. It relies on two sources of data to measure and analyze the mobility of students over time: Census estimates of the number of children living in each town in Massachusetts in 1980 and 1990, and annual public school enrollments from 1980 to 1995.

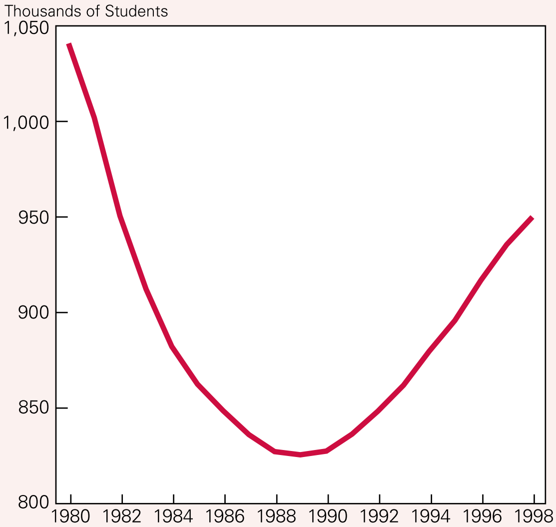
The study reports three major findings. (1) Net public school enrollment changes are positively related to differences across communities in school quality. (2) Shifts in enrollments were much more pronounced in the 1990s, when aggregate enrollments were rising and the economy was improving. (3) Proposition 2½ appears to have significantly altered the pattern of enrollment changes, with families with students moving to districts less constrained by this property tax limit.

Economic models predict that communities will adjust the amount of public services in response to the changing desires of households. However, in Massachusetts Proposition 2½ raised strong barriers to providing desired services in some communities.

The article is organized as follows. Section I documents the large cross-sectional differences in public school enrollment changes across Massachusetts communities in the 1980s and 1990s and relates them to differences in test scores. The next section discusses the manner in which households make residential location choices, sorting themselves among localities, and describes the economic, demographic, and political changes that affected these outcomes in the Commonwealth over the sample period. Section III presents regression results that examine the relationship between various community characteristics related to this broad context and the difference between actual

Figure 1

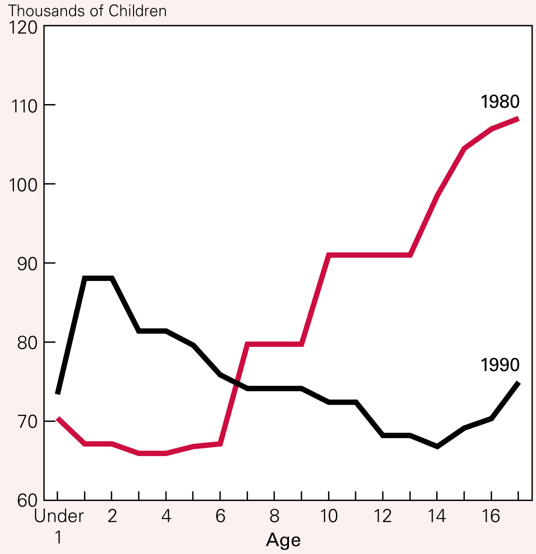
Elementary and Secondary Public School Enrollment in Massachusetts



Note: Enrollment as of October 1 of previous year.
Source: Massachusetts Department of Education, Individual School Report.

Figure 2

Age Distribution of Children in Massachusetts
1980 Census and 1990 Census



Note: Age distribution is interpolated for Census age groups that include more than one year (1–2, 3–4, 7–9, 10–11, 12–13).
Source: U.S. Bureau of the Census, Census of Population and Housing.

and demographically predicted changes in enrollments from 1980 to 1985 and 1990 to 1995. The conclusion explores the implications of the results.

I. Patterns of Enrollment Change in Massachusetts in the 1980s and 1990s

Public school enrollments declined statewide in Massachusetts in the 1980s and expanded in the 1990s (Figure 1). The 1980s decline was driven by demographics in that birth cohorts in the state had been shrinking over time. The situation reversed around 1990 when statewide enrollments began to expand. This change, too, was demographically driven, as birth cohorts grew larger beginning in the mid 1980s, leading to larger cohorts of first-graders replacing smaller cohorts of graduating high school seniors starting around 1990. This general pattern is consistent with the experience of other states over the same period. Figure 2 depicts the age distribution of preschool and school-age children in Massachusetts in 1980 and 1990. In 1980, the small preschool cohorts relative to large high school cohorts presage the ensuing enrollment losses. Similarly, in 1990 the bulge in preschool-age children compared to school-age children foretells the early 1990s enrollment gains.

Table 1 reports statewide enrollment in grades 1 through 8 from school year 1979–80 to school year 1994–95 (henceforth referred to, like fiscal years, as 1980 and 1995), with the number of public school students dropping in the early 1980s, stabilizing in the second half of the 1980s, and rising in the early 1990s.[1]

Since every student who would be in grade 1 and above in 1985 had been born by 1980 (and similarly for 1995 by 1990), Figure 2's data on the distribution of

[1] Grades 1 to 8 are used for this exercise rather than the full range of kindergarten through twelfth grade because there is more "noise" in the data for the highest and lowest grades. Kindergarten attendance is not required by Massachusetts law; hence some of the variation among communities in kindergarten enrollment simply reflects parental decisions about children's "readiness" for school and choices between day care and school enrollment. Similarly, dropouts become a factor in high school, introducing another difference among communities in public school enrollment that is unrelated to parental decisions about residential location or private school. In addition, regional vocational school enrollments are more difficult to allocate to individual cities and towns than other regional school enrollments; vocational schools operate only in grades 9 and up.

Table 1
Public School Enrollment in Massachusetts, 1980 to 1995
Grades 1 through 8, Statewide

	1980	1985	1990	1995
Enrollment (000)	608.2	491.6	502.6	560.4
5-Year % Change		−19.2		11.5
Population (000)				
Age 6 to 13	670.4	n.a.	579.7	n.a.
Age 1 to 8	559.7	n.a.	642.9	n.a.
Predicted 5-Year % Change in Enrollment[a]		−16.5		10.9
Actual minus Predicted (percentage points)		−2.7		.6

n.a. = not available
[a]Predicted change in enrollment is the beginning-of-period ratio of (population age 1 to 8) to (population age 6 to 13) minus 1, expressed as percent.
Source: Massachusetts Department of Education, U.S. Bureau of the Census.

preschool and school-age population of Massachusetts residents from the decennial Census provide a quantitative demographic "prediction" of enrollment changes in the first half of each decade.[2] These predictions indicate how enrollment would change if no families moved into or out of the state and private school enrollment rates were unchanged.[3]

For the state as a whole, public school enrollment in grades 1 to 8 shrank 19.2 percent from 1980 to 1985; the 1980 age mix implied a 16.5 percent decline in the ensuing five years. Presumably some families with school-age children moved out of the state or out of the public schools in the early 1980s, making the actual loss somewhat greater than the predicted loss. Net migration was slightly positive for Massachusetts during the 1980–85 period, but net migration results from large gross flows in each direction. Families with school-age children may have moved away, on net, and been (more than) replaced by other family types moving in; private school enrollment rates did rise in the 1980s (no data are available for the mid-decade year 1985).

Between 1990 and 1995, enrollments rose 11.5 percent. The 1990 Census showed a cohort aged 1 to 8 that was 10.9 percent bigger than the overlapping cohort aged 6 to 13, predicting a gain only slightly smaller than what actually occurred. Again, some combination of net in-migration of families with school-age children and declining private school enrollment rates could explain the small gap between the predicted and actual changes. Overall, Massachusetts experienced substantial out-migration of population, on net, in the 1990–95 period. But it may be that the out-migration response to the severe downturn of the early 1990s was concentrated among households without children, and that families with children were moving in and were somewhat more likely to attend public schools. In addition, the recession may have caused some resident families to forgo private schools for their children.

While demographic swings reduced and then augmented public school enrollments statewide, enrollments in individual communities were not moving in lockstep. Individual districts experienced these statewide shifts to a greater or lesser degree, depending not only on their beginning-of-period age mix but also on net movements of families with children into or out of the district and into or out of private schools in each period. Table 2 reports average actual and predicted 1980–85 changes in enrollment for communities grouped by local public school quality as measured by standardized test scores. (See the Box for a description of this measure of school quality and Appendix Table A1 for a list of cities and towns in each group.) In the early 1980s when enrollments were dropping statewide, enrollments in higher-quality school districts fell faster than enrollments in lower-quality districts, on average (column 1). The top 5 percent of communities lost an average of 22 percent of their students from 1980 to 1985 while the lowest 5 percent lost 14 percent.[4]

These differences among communities in rates of enrollment growth are consistent with what would be predicted by demographics, but movements among communities offset some of the effects of differences in 1980 age mix. According to the 1980 Census, the

[2] Because Census data are collected only every 10 years, this calculation can be done only for the first half of each decade.

[3] Private school enrollment rates are implicit in the Census-year ratio of public school enrollment to population. The predictions also implicitly assume that death rates are the same for preschool and school-age children. Since death rates are, in fact, higher among younger children, predicted enrollment growth is overestimated.

[4] The "all communities" differences between actual and predicted enrollment changes in Tables 2 and 3 do not exactly match the aggregate differences shown in Table 1 for two reasons: (i) Tables 2 and 3 report averages for only 321 communities (30 communities with 1980 enrollments under 150 excluded) while Table 1 includes all 351 cities and towns in the state; (ii) data observations for each community are weighted by beginning-of-period enrollment in calculating average changes by group in Tables 2 and 3 whereas Table 1's statewide totals implicitly weight each change by its own denominator.

Table 2
1980–85 Enrollment Changes by School Quality
Average percent change in grades 1–8 enrollment for 321 Massachusetts communities grouped by school quality

School Quality Rank	Actual Enrollment Change	Predicted Enrollment Change	Difference (percentage points)	Number of Communities
Lowest 5 percent	−13.5	−9.0	−4.5	16
Next 20 percent	−20.5	−14.1	−6.4	65
Below-median quartile	−20.8	−18.4	−2.4	79
Above-median quartile	−20.2	−19.5	−.7	80
Next 20 percent	−22.9	−24.7	1.7	66
Highest 5 percent	−22.4	−27.9	5.5	15
All	−19.2	−16.8	−2.5	321

Note: See Table 1 for explanation of predicted enrollment change.
Community data weighted by 1980 enrollment in grades 1 through 8.
School quality measure is average eighth-grade math and reading test scores, 1988-94.
See Appendix Table A2 for variable definitions and sources.

Table 3
1990–95 Enrollment Changes by School Quality
Average percent change in grades 1–8 enrollment for 321 Massachusetts communities grouped by school quality

School Quality Rank	Actual Enrollment Change	Predicted Enrollment Change	Difference (percentage points)	Number of Communities
Lowest 5 percent	5.3	17.8	−12.6	16
Next 20 percent	11.7	12.3	−.6	65
Below-median quartile	11.4	6.9	4.4	79
Above-median quartile	12.4	7.5	4.9	80
Next 20 percent	19.0	5.3	13.7	66
Highest 5 percent	22.5	3.7	18.8	15
All	11.5	10.7	.8	321

Community data weighted by 1990 enrollment in grades 1 through 8.
Note: See Table 2.

high-test-score communities had relatively fewer young children than communities with lower school quality, leading to a prediction of greater enrollment declines (in percentage terms) in high-score communities (column 2). The third column shows the difference between actual and predicted enrollment changes, which is a measure of intercommunity migration and shifts in private school attendance over the five years. These data indicate that migration and changes in private school attendance rates augmented public school enrollments in communities with higher-quality schools, compared with predictions, between 1980 and 1985. Many families with young children in 1980 apparently moved out of low-score communities as their children grew older, adding to public school enrollments in towns with higher test scores.

Table 3 reports similar data for the 1990 to 1995 period. The difference between actual and predicted enrollments across communities is much larger in this time period than the previous one. As enrollments rose statewide, high-score districts enjoyed greater student growth than low-score districts, with first through eighth grade headcounts rising 23 percent in the 15 towns with the highest test scores and only 5 percent in the 16 communities with the lowest scores. The age mix in 1990 predicts the opposite pattern of enrollment changes; as in 1980, the high-test-score communities had fewer preschool and primary-grade children relative to those of school age than the lower-score communities. Interjurisdictional movements and changes in private school enrollment patterns more than offset the initial demographics. These movements reduced the relative growth in student counts in the communities with the lowest test scores by over 12 percentage points, on average, and added 19 percentage points to the gains of the highest test-score communities.

Thus, net intercommunity movements of families with children and changes in private school attendance rates were positively associated with school quality during both periods, 1980–85 and 1990–95. The attraction of high-quality schools was much more pronounced in the 1990s, when enrollments were generally rising, than in the 1980s, when decreasing numbers of youngsters arrived at the schoolhouse door each year.[5]

[5] The school quality index is actually measured in the years surrounding 1990, as noted in the Box. It may be that the weaker relationship between school quality and net enrollment change in the 1980s as compared with the 1990s is partly due to a weaker relationship between the school quality index and actual school quality in the early 1980s (as observed or perceived by parents).

Measure of School Quality

Local school quality is measured by the average score on eighth grade reading and math assessment tests between 1988 and 1994. Community rankings differ very little on reading versus math, so the two are summed in this measure. The Massachusetts Educational Assessment Program (MEAP) tests were given to all Massachusetts public school eighth-graders in even-numbered years, but they have been discontinued in favor of a new set of tests beginning in 1998. While these tests are useful in generating a relative ranking of school quality across districts, they are not designed for comparison between years. In fact, the mean and variance of the scores are standardized in each year and the Massachusetts Department of Education explicitly warns against using the tests for comparisons over time. Consequently, the average score for each public school district, averaged over the years 1988 through 1994, is used only as a measure of cross-sectional differences in quality.

Test scores reflect a specific aspect of school quality—how well the average student performs on standardized tests. Test scores vary across districts both because the quality and quantity of "school inputs" such as teachers, books, and science labs vary across districts and because the quality of "student inputs" varies—some students enter school with more barriers to learning than others. For example, physical and mental disabilities and poor nutrition may handicap individual students; so may a lack of books or role models for academic achievement at home. Test scores can also be influenced by "peer effects," a term that encompasses the role of a student's neighborhood and in-school classmates in shaping expectations and performance. Because many of these influences on test scores cannot be determined by educational or budget decisions of schools or districts, test scores are a poor measure of "school quality" in the sense of how well a school is able to advance the learning of the students who attend it. But from the point of view of parents who want "the best education" for their children, peer effects can be just as important as school-determined factors such as teacher quality and curriculum. Thus, test scores can be a useful shorthand measure of the broader aspects of school quality that matter to parents choosing public schools for their children.

II. The Demographic, Economic, and Fiscal Context in Massachusetts

Differences in the degree of sorting between the 1980s and 1990s may be explained by a number of factors. Households' residential location choices will be influenced by the availability of public services and other amenities in individual cities and towns if they value these community characteristics to different degrees and if they are relatively mobile across communities. Any household's demand for housing in a particular community will reflect the degree to which its own preferences match the town's characteristics, as well as its ability to pay relative to the cost of locating there.[6]

At any point in time, households would be expected to have sorted themselves among communities according to their valuation of the specific attributes available in each place. For example, if a particular characteristic such as school quality is more important to one group of households, such as those with children, then one would expect to see more households with children in communities with high-quality schools, unless other types of households are able to outbid them for some reason. This sorting process is ongoing, as new households continuously enter the market. Furthermore, existing households might alter their location choices as their own preferences change, as community attributes change, or as the sorting and hence demand for housing of other households is disturbed by other shifts.

Thus, the redistribution of students across towns results from the interaction of demographics, economic conditions, and the housing market with local public sector attributes such as school quality. These factors form the context in which residential mobility occurs, and should help explain why actual public school enrollments in both 1985 and 1995 were greater

[6] Ross and Yinger, forthcoming, summarize a broad literature related to the allocation of households among communities, focusing on the role and endogeneity of the local public sector.

Table 4
Homeownership by Age of Head of Household

Age of Head	U.S. Homeownership Rate (Percent) 1980	1990
All	65.6	64.1
<25	21.3	15.3
25–29	43.3	35.9
30–34	61.1	51.5
35–39	70.8	63.1
40–44	74.2	70.4
45–54	77.7	76.1
55–64	79.3	80.4
65–74	75.2	78.7
>74	67.8	71.0

Source: Joint Center for Housing Studies, Harvard University, "The State of the Nation's Housing," 1991.

Table 5
Presence of Children by Age of Head of Household

Age of Head	U.S. Households with Children as Percent of all Households 1980	1990
All	39.1	33.9
<25	33.8	31.1
25–34	60.0	54.6
35–44	74.8	64.1
45–64	28.9	32.6
>64	2.8	3.1

Source: Joint Center for Housing Studies, Harvard University, "The State of the Nation's Housing," 1991 and 1996.

in communities with better schools than what would have been predicted assuming no households with children had moved into or out of any town. Furthermore, changes in this context may explain why the flight to quality was especially pronounced in the 1990s.

Demographic Bulge, Household Incomes, and the Housing Market

Aggregate enrollments were declining in the 1980s and rising during the 1990s, a sea change that one would expect to be associated with a shift in the strength of overall demand for and interest in schools and school quality in different locations. "Baby boomers," born during the period of high birth rates between 1946 and 1964, produced a bulge in the population distribution that has had a substantial effect on many areas of the economy.[7] The size and shape of the baby-boom bulge was similar in Massachusetts and the nation, although the post-boom cohort—the "baby bust"—was somewhat smaller in Massachusetts; that is, the post-boom drop-off was somewhat steeper in Massachusetts.

In 1980, the boomers were between the ages of 16 and 34. The bulk of them were in their twenties and, as shown in Table 4, home ownership rates have always been low for this age group. In 1980, only 21 percent of U.S. households with the head under age 25 and 43 percent of households headed by individuals between 25 and 29 years old owned homes, while over three-fifths of older householders were homeowners. Similarly, households with younger heads are less likely to contain children. Table 5 reports the fraction of households with children by age of head. In 1980, only one-third of households with head under age 25 included children, as compared with three-fifths of those with heads ages 25 to 34 and three-quarters of those in which the head was between 35 and 44 years old.

As the baby boom cohort aged from 1980 to 1985 (when they ranged in age from 21 to 39), an increasing number of its members had children and many young baby boomers entered the housing market as first-time buyers. Not only were these households moving up their "age-earnings" profiles, but per capita income was rising faster in Massachusetts than in the United States as a whole (see Case 1986). Housing prices began to rise sharply in late 1984, but did not explode upwards until 1985. (See Figure 3 and Case and Mayer 1996).

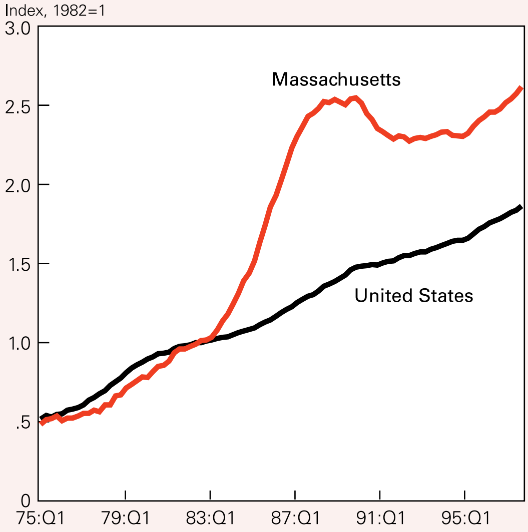
The period between 1985 and 1990 was one of dramatic change. Home prices in the average Massachusetts community nearly tripled. A home purchased for $100,000 in 1982 sold for $275,000 in 1989 at the peak of the market, with most of the appreciation

[7] In a provocative and often cited paper, Mankiw and Weil (1989) argue that the aging of the "baby boom" would lead to a substantial decline in housing prices as that cohort aged and moved out of the housing market. While its drastic conclusions have been challenged by a number of authors, it highlights clearly the potential importance of demographics. In a related study, Case and Mayer (1996) show that differences in demographics due to the baby boom can affect the relative price of housing across different communities.

Figure 3

Home Prices

Repeat Sales Price Index for Homes with Conventional Mortgages



Source: Fannie Mae and Freddie Mac.

Table 6

Equity in 1990 in a $100,000 Home by Year of Purchase

Year Purchased	Equity in 1990
1982	$170,000
1983	146,000
1984	106,000
1985	69,000
1986	40,000
1987	24,000

Note: Includes appreciation based on Massachusetts repeat sales conventional mortgage home price index and amortization of an 80 percent loan at 8 percent.

Source: Authors' calculations based on data from Fannie Mae and Freddie Mac.

occurring between 1985 and 1988 (Case and Mayer 1995). The largest price increases were recorded in the lower-income towns where many first-time buyers resided. Then in 1989, the Massachusetts "miracle" began to dissolve. The region's real estate boom turned to bust and the New England region headed into the 1990–91 recession well before the rest of the nation.

By 1990, the boomer cohort was between 26 and 44—mostly in its thirties. While the younger boomers were still first-time home buyers, the older group had children in school and were ready to trade up. Aggregate school enrollments were rising. The recession of 1990–91 was followed by recovery beginning in 1992, and conditions were ripe for older boomers with children to purchase homes in better school districts. Housing prices had come down between 10 and 15 percent and mortgage interest rates fell to record lows in 1992 and 1993. While incomes stagnated in 1990 and 1991, they resumed their upward trend in 1992 and, in addition, the boomers were moving into their peak earning years.

Equally important, households wanting to trade up came to the table with equity. Table 6 shows the equity buildup by mid 1990 for a household that had bought a $100,000 home between 1982 and 1987. The average household buying a home before 1986, as many baby-boomers did, had gained between $70,000 and $170,000 in equity by 1990. The 1989–92 fall in housing prices resulted in only a small decline in equity relative to the sizable gains of the mid 1980s.

The combination of cash buildup, low interest rates, and higher and rising incomes meant that trade-up buyers could now afford to buy into the school systems that they and their predecessors could not afford a decade earlier. And the aging of the baby boomers meant that more of them had children, implying that a greater fraction of households were interested in schools. As a result, the distribution of enrollments dramatically shifted towards the better school districts between 1990 and 1995.

Change in Local Fiscal "Rules"

Added to these demographic and economic changes in Massachusetts was the property tax limitation measure, Proposition 2½. Passed by voters in November 1980, it required communities in Massachusetts to reduce their property tax levies by 15 percent per year until they attained a maximum rate of 2.5 percent of the market value of property. Once those cuts had occurred, the property tax levy could not exceed the town's levy limit, which rose by only 2.5 percent per year, plus an allowance for new development, unless local voters passed an override to increase taxes more.

About two-fifths of the state's 351 cities and towns were required to make budget cuts in fiscal year 1982; all had reached the 2.5 percent rate limit by fiscal year 1984. After the initial cuts, Proposition 2½ had only minor effects on local budgets from 1985 to 1990 because of three favorable factors: The Commonwealth provided sizable increases in local aid, the real estate boom led to considerable new development that increased the tax base, and school enrollment declines reduced pressures on local budgets. However, these favorable trends reversed at the end of the decade; the local economy went into a nose-dive, the state cut back on aid, and the baby boom echo caused enrollments to pick up. As a result, an increasing number of communities bumped up against their levy limits.[8]

Table 7
1980–85 Enrollment Changes by Degree of Constraint
Average percent change in grades 1–8 enrollment for 321 communities grouped by number of years of initial property tax reductions

Proposition 2½ Constraint in 1980s	Actual Enrollment Change	Predicted Enrollment Change	Difference (percentage points)	Number of Communities
Zero years of cuts	−18.0	−19.7	1.7	166
One year of cuts	−21.6	−18.8	−2.8	122
Two or three years of cuts	−17.4	−11.6	−5.9	33
All	−19.2	−16.8	−2.5	321

Note: See Table 1 for explanation of predicted enrollment change.
Community data weighted by 1980 enrollment in grades 1 through 8.
See Appendix Table A2 for variable definitions and sources.

Table 8
1990–95 Enrollment Changes by Degree of Constraint
Average percent change in grades 1–8 enrollment for 321 communities grouped by leeway between levy and levy limit in fiscal year 1989

Proposition 2½ Constraint in 1990s	Actual Enrollment Change	Predicted Enrollment Change	Difference (percentage points)	Number of Communities
Not at levy limit	15.0	8.0	7.0	193
At levy limit	8.9	12.7	−3.8	128
All	11.5	10.7	.8	321

"At levy limit" defined as levy greater than or equal to 99.9 percent of levy limit.
Note: See Table 1 for explanation of predicted enrollment change.
Community data weighted by 1990 enrollment in grades 1 through 8.
See Appendix Table A2 for variable definitions and sources.

Thus, in the early 1990s, as baby-boomer families began to purchase trade-up homes, they may have looked not only at the school quality of each community in which they could locate but also at the degree to which each community was constrained by the growth-limit provisions of Proposition 2½. Communities required by Proposition 2½ to reduce their taxes might have been more attractive to families with children who preferred direct voter oversight of tax increases and who believed municipal officials were likely to overspend in the absence of such budget constraints. Alternatively, more constrained cities and towns might be less attractive because they were less able to respond to resident preferences regarding local public services, both school and nonschool.[9] Tables 7 and 8 suggest that the latter factors dominated, that is, families with children found constrained communities less attractive, on average. Communities facing greater Prop 2½ property tax reductions in the early 1980s saw bigger declines in enrollment between 1980 and 1985, net of initial demographics, than communities facing no cuts. And between 1990 and 1995, communities bumping against their levy limits before the period began (fiscal year 1989) similarly saw smaller enrollment increases than less constrained communities.

However, just as for school quality, the observed pattern of enrollment changes may result from other characteristics of these communities. That is, Proposition 2½ may have disproportionately affected communities that might otherwise have failed to attract public school children. To look at how the variety of eco-

[8] The number of cities and towns whose property tax levies were within 0.1 percent of their levy limits jumped from 82 in FY1988 to 224 in FY1990.

[9] In an earlier paper, Bradbury, Mayer, and Case (1997) found that Proposition 2½ significantly reduced school and nonschool spending in constrained communities between 1990 and 1994, and thereby reduced house prices in constrained communities, other things equal.

nomic, demographic, and fiscal forces described above influenced families with school-age children as they sorted themselves out among communities and made choices about public versus private schooling, one needs more than the single-variable cross-tabulations reported in Tables 2, 3, 7, and 8. The regressions explored in the next section of the paper explain community enrollment changes, controlling for beginning-of-period demographics, and allow quantification of the magnitude of some of these influences.

III. Economic, Demographic, and Fiscal Influences on Enrollment Changes, 1980–85 and 1990–95

As the discussion in the preceding section makes clear, many factors would be expected to influence families' residential location choices, and the importance of various factors might differ for families with children compared with childless households. To examine the attractiveness of specific community characteristics to families with children, regressions are estimated to explain net enrollment changes in grades 1 through 8 between 1980 and 1985 and from 1990 to 1995. Net enrollment change, the dependent variable, is defined as the difference between the actual percentage change in enrollment in grades 1 through 8 and the predicted enrollment change, where the predictions are calculated as the percentage difference between the number of resident children ages 1 to 8 and the number ages 6 to 13 at the beginning of the period (1980 or 1990).[10] Net changes over each five-year period reflect movements of families with school-age children among communities and shifts in private school enrollment patterns.[11] The equations are estimated across 321 cities and towns in Massachusetts; 30 of the state's 351 communities with total 1980 enrollments under 150 have been excluded.[12] Appendix Table A2 reports variable means and sources.

Basic Influences: Local Public Sector Attributes and Developable Space

Families with school-age children are likely to care more about school quality than families with no children. Whether enrollments are generally declining or rising, net enrollment changes are likely to be greater (more positive) in higher-quality districts, other things equal, as mobile families attempt to move their children into the better public schools.

Whether the constraints imposed by Proposition 2½ would have differential effects on households with children as compared with childless households is less clear. Nonetheless, like any other factor that alters the attractiveness of individual communities, Prop 2½ constraints will affect public school enrollments indirectly as families make their locational choices on a variety of grounds, and then send their children to local public schools. As discussed in the previous section, cities and towns forced into a cutting mode early on and those that faced tighter restrictions going into the 1990s might be more attractive if Proposition 2½ provided a needed restraint and direct voter control of tax increases that were positively valued by residents, or they might be less attractive to potential residents than communities that were able to make spending decisions independent of the levy limits imposed by Proposition 2½.[13]

Aside from these public sector attributes, the potential for growth in the local housing stock can be expected to influence gross enrollment changes. Communities with more developable land could accommodate more additional households (with or without children) in any five-year period, via new construction, than communities that were already more densely developed.

Columns 1 through 3 of Table 9 report estimates from a sparse regression that includes the public sector attributes and availability of developable land. The school quality results are roughly consistent with the patterns shown in Tables 2 and 3, indicating a positive association between school quality and net enrollment increases in both time periods. As in the simple cross-tabs, the relationship is much stronger in the 1990s. According to the 1990–95 coefficient estimate, a community with test scores 140 points (one standard deviation) above the mean would see about 6 percentage points faster growth in enrollments between 1990 and 1995 than a community with average

[10] This is the "difference" variable shown in column 3 of Tables 2, 3, 7, and 8.

[11] The measure is labeled "net" change because it nets out initial differences in age mix and private school enrollment rates.

[12] For this cutoff, total enrollment is defined to include all grades—pre-kindergarten through grades 12, 13, 14, and ungraded.

[13] The degree of constraint imposed by Prop 2½ may be seen by potential residents as an indication of future changes in school quality. David Figlio (1998) finds that the property tax limitation Measure 5 significantly increased student-teacher ratios in Oregon's school districts, and the effect was more pronounced in districts more constrained by the measure (those that relied more heavily on local property tax revenues before Measure 5). Also see footnote 9 above.

Table 9
Regression Results
Net Changes in Grades 1 to 8 Enrollment, 1980–85 and 1990–95
Dependent Variable: Percentage Point Difference between Actual and Predicted Enrollment Change
(Robust standard errors in parentheses below estimated coefficients)

Independent Variables:	1980–85 (1)	1990–95 (2)	1990–95 (3)	1980–85 (4)	1990–95 (5)	1980–85 (6)	1990–95 (7)
School quality index	7.36*	41.4***	44.6***	.0654	31.6***	.425	29.7***
	(4.10)	(5.7)	(4.7)	(6.78)	(7.2)	(6.69)	(7.2)
Prop 2-1/2 initial revenue cuts for one year	−1.03	−1.11		−.406		−.664	
	(1.19)	(1.67)		(1.21)		(1.17)	
Prop 2-1/2 initial revenue cuts for two or three years	−5.63***	−4.02		−3.71*		−3.86**	
	(1.88)	(2.76)		(1.93)		(1.86)	
At levy limit in FY1989		−3.17**	−3.60**		−3.16**		−2.89**
		(1.52)	(1.42)		(1.30)		(1.33)
Developable land per housing unit (1984)	2.05*	1.09	1.42*	1.73	.833	1.97**	1.05*
	(1.18)	(.824)	(.78)	(1.14)	(.627)	(.98)	(.61)
Median family income ($000) in 1980 or 1990				.400**	.453***	.599***	.526***
				(.189)	(.107)	(.219)	(.117)
% of resident adults college-educated in 1980 or 1990				.0437	−.109	.196*	−.0124
				(.0836)	(.101)	(.104)	(.122)
Dummy variable: Located inside Boston metropolitan area				−5.17***	−8.30***	−4.36***	−7.12***
				(1.26)	(1.91)	(1.26)	(1.99)
Median house value ($000) in 1980 or 1990						−.222***	−.0210
						(.080)	(.0309)
Median unit rent ($) in 1980 or 1990						.0162	−.0124*
						(.0130)	(.0065)
Constant	−20.7*	−104***	−14***	−8.25	−91.5***	−9.55	−82.0***
	(11.2)	(15.8)	(13)	(16.2)	(17.2)	(16.3)	(17.5)
R-squared	.107	.251	.246	.151	.322	.178	.331
Number of observations	321	321	321	321	321	321	321

***significantly different from zero with 99% confidence or better.
**significantly different from zero with 95 to 99% confidence.
*significantly different from zero with 90 to 95% confidence.
See Appendix Table A2 for variable definitions.

scores, other things equal. By the same token, two communities with scores differing by 625 points—the difference between the average community in the top 5 percent of scores and the average in the bottom 5 percent—would show almost 28 percentage points difference in net enrollment change, equal to the bulk of the unconditional difference shown in Table 3.[14]

Several explanations are possible for the stronger effect of school quality in the 1990s. As discussed earlier, the differential effects in the two periods may be related to aggregate enrollment trends. When statewide counts of school-age children are rising, a greater fraction of families will include children and hence care about school quality, and families with increased numbers of children will care more strongly about school quality. The combination of this demographic swing and shifts in family income and house prices is likely to have altered the observed net mobility responses to local school quality. An additional factor relates to national trends that altered the payoff to education. During the 1980s, the distribution of earn-

[14] The figures shown in Table 3 are "unconditional" in the sense that they control for no other variables; the regression coefficient in column 3 controls for associated variations in the other included variables.

ings became more unequal as less educated workers saw their real earnings fall markedly while the labor market returns to higher education rose. As parents became increasingly aware of the importance of education to their children's future prospects, they may have placed a heavier weight on the quality of local public schools in making their family's residential location choices.

During the 1980s less educated workers saw their real earnings fall markedly, and parents became increasingly aware of the importance of education to their children's future prospects.

Proposition 2½ appears to have affected enrollments as well. Variables measuring the number of years of initial budget cuts that the tax limitation required are negatively associated with net enrollment growth from 1980 to 1985. Specifically, communities required to cut their budgets for two or three years saw almost 6 percentage points greater enrollment losses over the five years than communities with no initial cuts mandated, other things equal, although communities with one year of cuts were indistinguishable (in terms of 1980–85 net enrollment changes) from those with no cuts.[15]

These measures of initial cuts are unrelated to net enrollment changes a decade later (column 2). But Proposition 2½'s limits on revenue growth affected increasing numbers of communities toward the end of the 1980s, and a variable measuring the stringency of the growth limit does help to explain enrollment changes from 1990 to 1995.[16] Specifically, communities at their levy limits in 1989 experienced smaller enrollment increases between 1990 and 1995, 3 to 4 percentage points less, than did communities less constrained by Proposition 2½.[17] Whether by cutting the quality of schools in ways not captured by the test score measure, reducing the quality of nonschool public services, or inhibiting community flexibility more generally, Proposition 2½ made constrained communities relatively less attractive to families with children, both in the early 1980s and the early 1990s.[18]

Land use data for Massachusetts communities are available only for 1984. A 1984 measure of open and residential land relative to the 1980 housing stock ("developable" land) was associated with net enrollment growth in both periods.[19] Several explanations are possible for the finding of a larger estimated effect of developable land in the 1980s. First, space to accommodate added families was probably more at issue in the early 1980s than in the early 1990s. Two pieces of evidence suggest that the number of households was expanding more in the earlier period despite the fact that enrollments were shrinking: (1) The state experienced net in-migration between 1980 and 1985 while population moved out, on net, in the early 1990s. (2) Based on permits, more housing was constructed in Massachusetts in the 1980–85 period than in the 1990–95 period. A second explanation has to do with measurement: The measure of developable land

Proposition 2½ made constrained communities relatively less attractive to families with children, both in the early 1980s and the early 1990s.

[15] The unconditional difference in net enrollment growth between communities with zero years of cuts and those with two or three years of cuts (shown in Table 7) is 7.6 percentage points.

[16] Because column 2 indicates the initial FY1982–84 property tax cuts had no discernible effects on enrollment growth by the early 1990s once the current (FY1989) constraint imposed by Proposition 2½'s growth limits is controlled for, column 3 reports a reestimate of the 1990–95 equation dropping the two variables measuring years of initial cuts. The other coefficient estimates are changed very little. This pared-down version is carried forward into columns 5 and 7 as additional variables are included.

[17] The unconditional difference shown in Table 8 is 10.8 percentage points. Thus, differences between constrained and unconstrained communities in the other included variables (school quality, developable land) account for more than half of the observed difference in net enrollment change from 1990 to 1995.

[18] Among communities at their levy limits in FY1989, one might expect those that had passed at least one override to be less constrained than communities that had never passed an override. Override passage prior to FY1990, however, is not associated with 1990–95 enrollment changes in a statistically significant manner when "at levy limit" is controlled for.

[19] Ideally, the quantity of developable land would be adjusted for local zoning restrictions, but such data are not available.

may relate more strongly to growth in the 1980s because the year land use was tallied, 1984, obviously precedes the 1990–95 period by six-plus years. To the degree that the mid-decade 1980s housing boom elicited residential development on open land, this 1984 measure may not reliably represent differences among communities in the availability of land as the 1990–95 period opened.[20]

Other Local Amenities

Because of peer effects in local schools as well as more general neighborhood effects, families with children might be attracted to communities whose residents have higher incomes or are more highly educated. Location is another key local characteristic; proximity to jobs, retail services, and other economic activity would be viewed as a plus by most households, although some household types may be willing to pay more for increased access than others.

In both periods, communities with higher-income residents at the beginning of the period realized greater net enrollment gains. Communities near Boston experienced larger net enrollment declines in the 1980s and smaller increases in the 1990s than communities outside the Boston area.

The regressions reported in columns 4 and 5 of Table 9 add measures of these other local amenities to the analysis. In both periods, communities with higher-income residents at the beginning of the period realized greater net enrollment gains. By contrast, the percentage of residents with a college or higher education is unrelated to enrollment changes.[21] An indicator of geographic location is also included in the equations. Coefficient estimates on a dummy variable for the Boston metropolitan area imply that communities near Boston experienced larger net enrollment declines in the 1980s and smaller increases in the 1990s than communities outside the Boston area.[22] The effect of location was somewhat stronger in the 1990s than in the 1980s.

Inclusion of this location variable reduces the importance of developable land in both periods, especially 1990–95. This effect is understandable, as the two variables are correlated (developable land is more abundant outside the metro area of the state's largest city) and likely to be picking up the same effect —more room for additions to the housing stock (and hence enrollment growth) at greater distances. Since one would expect metro Boston location to be an attraction, not a disamenity, the "room to grow" effect must be dominating the accessibility effect. Put another way, households without public school children must be willing to pay more for access to Boston's concentration of economic activity, outbidding households with public school children for these locations, other things equal.

Inclusion of the nonpublic sector amenity variables also reduces the size or statistical significance of other variables. Most notable is the 1980–85 coefficient on school quality, which becomes insignificantly different from zero once these other variables are included. These results suggest that the pattern in Table 2 and the coefficient estimate in column 1 of Table 9 may reflect some other factor such as median family income, percent of residents college-educated, or location that is associated with both school quality and net enrollment change. Alternatively, these variables might also be correlated with school quality as measured by student test scores. Students from more highly educated families are likely to have higher test scores independent of the schools that they attend. If such parents also choose to live in higher-quality school districts, it may be difficult to pick up an independent effect of school quality in this regression.

[20] The fact that 1984 is near the end of the 1980–85 period should bias the 1980–85 coefficient estimate downward, if anything, since development occurring between 1980 and 1983 would presumably be positively associated with added families but would reduce open land in 1984, other things equal.

[21] One might also expect families with children to be more sensitive to the (negative) amenity of crime than households without children. However, alternative versions of the regressions shown in columns 4 and 5 that also include a beginning-of-period crime rate obtain coefficient estimates on the crime rate that are not significantly different from zero.

[22] The Boston PMSA, as defined by the Census Bureau based on 1980 commuting patterns, includes 106 cities and towns in eastern Massachusetts.

Trade-off: Cost Differentials

As households consider the relative attractiveness of the "packages" of characteristics available in individual cities and towns, they must weigh their own preferences against the price they must pay to buy each package. The market-wide valuation of each community's traits is reflected in existing house prices and rents. Other things (all the attributes in the package) equal, communities with higher house prices or rents would be less attractive from a cost or affordability perspective.

> *The market-wide valuation of each community's traits is reflected in existing house prices and rents. Households must weigh their own preferences against the price they must pay.*

Columns 6 and 7 report coefficient estimates from equations that include the house price and rent variables—each community's median house value and median rent at the beginning of the period. As would be expected for cost measures, higher-priced communities generally saw smaller net enrollment gains or greater net losses, other things equal, in both periods. The effect of median house value is significantly different from zero only in the 1980s[23] and the effect of median rent only in the 1990s.

Inclusion of the housing cost variables alters several other coefficient estimates. Median family income and percent college-educated both have more positive effects on enrollment growth once house prices are controlled for. Because these characteristics are positive attractions, their historical effects on housing demand are capitalized in higher beginning-of-period house prices. Thus, the coefficients on the amenity variables in columns 4 and 5 reflect the net effect of the attraction of these attributes offset somewhat by the associated higher house prices. In columns 6 and 7, these cost effects are estimated separately, allowing the amenity coefficients to be just that. Most notable among these effects, the fraction of adult population with a college degree or more obtains a coefficient that is significantly different from zero in the 1980–85 period. The average community contained about 20 percent college-educated adults in 1980, with a range across the 321 communities from 5 to 60 percent. A community with 10 percentage points (a little less than one standard deviation) more college-educated residents in 1980 would have seen about 2 percentage points more enrollment growth over the 1980–85 period than a community with average educational attainment, other things equal.

Overall, the regression results indicate that differences in school quality and the stringency of Proposition 2½ altered rates of enrollment growth in Massachusetts communities during the early 1980s and early 1990s. Aside from these effects of the local public sector, a boom-bust-recovery economic cycle and the demographic phenomenon of the baby boom played out to influence enrollment changes differentially in the two periods. Enrollments generally declined in the early 1980s as the youngest baby boomers graduated from high school and left smaller cohorts behind. During this period, although school quality had no apparent effect, families with children were more likely to move into high-income communities outside the Boston metro area, with more developable land, more college-educated residents, and lower housing costs. In the early 1990s, by contrast, enrollments generally expanded as increasing numbers of children of baby boom parents entered their school-attending years. As family incomes and housing equity recovered from the bust, older baby boom families were able to trade up while younger baby boom families

> *Overall, differences in school quality and the stringency of Proposition 2½ altered rates of enrollment growth in Massachusetts communities during the early 1980s and early 1990s.*

[23] One explanation for the smaller effect of house prices in the 1990s is that more families were "trading up" rather than buying for the first time in the 1990s; the oldest baby boomers were only 34 in 1980 but 44 by 1990. For existing homeowners contemplating a move, higher house prices represent an increment to their buying power as well as the cost of what they intend to buy.

entered the market for the first time, facing housing costs that had come down from their speculative highs. The draw of higher-income neighbors continued from the 1980s, while the attraction of higher school quality became dominant. The increased numbers of households with the wherewithal to choose may have combined with growing recognition of the importance of education, raising enrollments in communities with higher test scores, on net.

IV. Conclusion and Outlook

The findings described here suggest that economic and demographic factors, along with the passage of Proposition 2½, have had a significant impact on reallocating public school enrollments across districts in Massachusetts. The study reports two major findings: First, the regression estimates support the hypothesis that school quality was a key determinant of family location decisions, and hence enrollment growth, in the early 1990s, when aggregate enrollments were rising and the economy was improving, but the effect of school quality was negligible in the 1980s once other factors were controlled for. Second, Proposition 2½ appears to have significantly shifted the pattern of enrollment changes, with students moving, on net, to districts that are less constrained by this tax limit. Compared with otherwise similar communities, cities and towns with two or three years of property tax cuts in the early 1980s saw greater enrollment declines from 1980 to 1985 and communities at their levy limits in 1989 saw smaller enrollment gains from 1990 to 1995.[24]

The Proposition 2½ results of this study are troubling because they suggest that the tax limitation may be interfering with the efficient sorting of families among communities.

A range of theoretical and empirical research suggests that some efficiency gains result from community specialization in the provision of local public services and the resulting sorting of residents among local jurisdictions according to their preferences and ability to pay. The school quality results are supportive of these hypotheses, as they imply that families made location choices that increased the number of children in high-quality schools. By contrast, the Proposition 2½ results are troubling because they suggest that the tax limitation may be interfering with that "efficient" sorting. Families with children appear to be "voting with their feet," moving out of communities that have run up against their tax limits and chasing communities that have excess capacity to support schools because they are below their mandated tax limit.

The attraction of school quality to families with children is unlikely to diminish, as educational attainment remains a key determinant of individual economic success.

These findings are consistent with previous work that examined the relationship between house prices and the constraints of Proposition 2½ (Bradbury, Mayer, and Case 1997). That research found house prices declining more or rising less between 1990 and 1994 in communities more constrained by Proposition 2½. The current paper addresses the "quantity" rather than "price" side of the housing market, focusing on families with children, and finds the (quantity) demand for housing lower in communities more constrained by Proposition 2½.

Looking forward, the importance of Proposition 2½ may diminish somewhat in the next few years, but the key role of school quality is likely to persist. The constraints of Proposition 2½ appear less binding as the economy improves and communities pass more overrides; the fraction of cities and towns "at" their levy limits has declined gradually since peaking in 1991. By contrast, the attraction of school quality to families with children is unlikely to diminish in the

[24] Because the binding provisions of Proposition 2½ shifted from the early 1980s (initial cuts) to the 1990s (growth limits), the measure of constraint, but not Prop 2½'s negative effect on enrollments, changed between the 1980s and the 1990s.

remainder of this decade or early in the next. For one thing, educational attainment remains a key determinant of individual economic success as income inequality remains high, and recent research has stressed the role of elementary and secondary education along with college.[25] In addition, the scarcity element of school quality associated with demographic pressures will continue to increase into the next decade, as enrollments continue rising.[26] Furthermore, new "high stakes" student testing and reporting mandated by Massachusetts' education reform law will focus additional attention on school quality and how it varies among communities.

Indeed, test results are expected to provide parents with information to hold local school districts "accountable" for school performance through city and town governmental channels as well as by voting with their feet as they have always done. A by-product of the reforms may thus be a broadening of the impetus to improve local education in below-average districts beyond the negative pressures that declining enrollments have historically exerted. To the degree that communities respond to this internal parental pressure, the need for interdistrict movements documented in this article should decline. In this context, however, Proposition 2½'s constraints may loom larger, as they impair some communities' access to the resources needed to upgrade local schools.

[25] For example, see McMurrer and Sawhill 1998, chapter 8.

[26] Based on the pattern of annual births in Massachusetts through 1996, the number of children in grades 1 through 8 will rise each year through 2001 and only then begin to ease slightly, while the number of children in first through twelfth grade will continue to rise.

References

Bradbury, Katharine L., Christopher J. Mayer, and Karl E. Case. 1997. "Property Tax Limits and Local Fiscal Behavior: Did Massachusetts Cities and Towns Spend Too Little on Town Services under Proposition 2½?" Federal Reserve Bank of Boston Working Paper 97–2, April.

Case, Karl E. 1986. "The Market for Single Family Homes in the Boston Area." *New England Economic Review*. May/June, pp. 38–48.

Case, Karl E. and Christopher J. Mayer. 1995. "The Housing Cycle in Eastern Massachusetts: Variations among Cities and Towns." *New England Economic Review,* March/April, pp. 24–40.

———. 1996. "Housing Price Dynamics Within a Metropolitan Area." *Regional Science and Urban Economics*, vol. 26, pp. 387–407.

Cutler, David M., Douglas W. Elmendorf, and Richard Zeckhauser. 1997. "Restraining the Leviathan: Property Tax Limitations in Massachusetts." Board of Governors of the Federal Reserve System Working Paper #1997–47.

Downes, Thomas A. and David Figlio. 1997. "School Finance Reforms, Tax Limits, and Student Performance: Do Reforms Level-Up or Dumb Down?" Tufts University, mimeo, June.

Figlio, David N. 1998. "Short-Term Effects of a 1990s-Era Property Tax Limit: Panel Evidence on Oregon's Measure 5." *National Tax Journal*, vol. LI, no. 1 (March), pp. 55–70.

Mankiw, N. Gregory and David N. Weil. 1989. "The Baby Boom, the Baby Bust, and the Housing Market." *Regional Science and Urban Economics*, vol. 19, no. 2 (May), pp. 235–58.

McMurrer, Daniel P. and Isabel V. Sawhill. 1998. *Getting Ahead: Economic and Social Mobility in America*. Washington, DC: Urban Institute Press.

Ross, Stephen and John Yinger. Forthcoming (1998). "Sorting and Voting: A Review of the Literature on Urban Public Finance." In Paul Cheshire and Edwin S. Mills, editors, *Handbook of Regional and Urban Economics, Volume 3, Applied Urban Economics*. Amsterdam: North-Holland.

Tiebout, Charles M. 1956. "A Pure Theory of Local Expenditures." *Journal of Political Economy*, vol. 90, no. 5, pp. 917–43.

White, H. 1980. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica*, vol. 50, pp. 483–99.

Appendix Table A1

Massachusetts Communities Ranked by Average Test Score 1988–94[a]

(Alphabetical within rank groups)

BOTTOM HALF

Lowest 5 percent
(scores 2220–2515)

- Boston
- Brockton
- Cambridge
- Chelsea
- Everett
- Fall River
- Fitchburg
- Holyoke
- Lawrence
- Lowell
- Lynn
- New Bedford
- Somerville
- Springfield
- Webster
- Worcester

Next 20 percent
(scores 2520–2630)

- Acushnet
- Amesbury
- Ashfield
- Athol
- Attleboro
- Ayer
- Belchertown
- Bellingham
- Blackstone
- Buckland
- Carver
- Charlemont
- Charlton
- Chicopee
- Colrain
- Conway
- Deerfield
- Dracut
- Dudley
- Fairhaven
- Freetown
- Gill
- Gloucester
- Granby
- Granville
- Halifax
- Haverhill
- Holbrook
- Hudson
- Kingston
- Lakeville
- Leominster
- Malden
- Mashpee
- Maynard
- Medford
- Methuen
- Middleborough
- Milford
- Millville
- Montague
- North Adams
- Northbridge
- Orange
- Oxford
- Pembroke
- Petersham
- Plymouth
- Plympton
- Quincy
- Randolph
- Revere
- Royalston
- Salem
- Shelburne
- Sunderland
- Taunton
- Tewksbury
- Waltham
- Ware
- Wareham
- Westport
- Whately
- Winchendon
- Woburn

Below median quartile
(scores 2631–2695)

- Adams
- Agawam
- Avon
- Barre
- Bernardston
- Blandford
- Bourne
- Bridgewater
- Brimfield
- Brookfield
- Cheshire
- Chester
- Clinton
- Danvers
- Dartmouth
- Dedham
- Douglas
- Easthampton
- Egremont
- Falmouth
- Franklin
- Gardner
- Groveland
- Hanson
- Hardwick
- Holland
- Hubbardston
- Hull
- Huntington
- Littleton
- Ludlow
- Mansfield
- Marion
- Marlborough
- Mattapoisett
- Merrimac
- Monson
- Montgomery
- Nantucket
- Newbury
- New Marlborough
- Northampton
- North Attleborough
- North Brookfield
- Northfield
- Norton
- Oakham
- Palmer
- Peabody
- Pittsfield
- Raynham
- Rochester
- Rockland
- Rowley
- Russell
- Salisbury
- Saugus
- Seekonk
- Sheffield
- Somerset
- Southbridge
- Southwick
- Stoneham
- Stoughton
- Sturbridge
- Sutton
- Swansea
- Uxbridge
- Wakefield
- Wales
- Warren
- Watertown
- West Brookfield
- West Newbury
- Weymouth
- Whitman
- Wilmington
- Winthrop
- Worthington

TOP HALF

Above-median quartile
(scores 2696–2790)

- Abington
- Arlington
- Ashburnham
- Ashby
- Ashland
- Auburn
- Barnstable
- Becket
- Berkley
- Berlin
- Beverly
- Billerica
- Boylston
- Braintree
- Burlington
- Canton
- Chesterfield
- Dalton
- Dennis
- Dighton
- East Bridgewater
- East Brookfield
- East Longmeadow
- Easton
- Edgartown
- Essex
- Foxborough
- Framingham
- Georgetown
- Goshen
- Grafton
- Great Barrington
- Greenfield
- Hampden
- Harwich
- Hinsdale
- Hopedale
- Ipswich
- Lancaster
- Lanesborough
- Lee
- Leicester
- Marshfield
- Melrose
- Millbury
- Millis
- Milton
- Nahant
- Newburyport
- Norfolk
- North Andover
- Norwood
- Oak Bluffs
- Pepperell
- Phillipston
- Plainville
- Provincetown
- Rehoboth
- Rockport
- Shirley
- Southampton
- South Hadley
- Spencer
- Stockbridge
- Templeton
- Tisbury
- Townsend
- Truro
- Tyngsborough
- West Bridgewater
- Westfield
- Westhampton
- Westminster
- West Springfield
- West Stockbridge
- West Tisbury
- Williamsburg
- Williamstown
- Wrentham
- Yarmouth

Next 20 percent
(scores 2791–2955)

- Amherst
- Bedford
- Belmont
- Boxford
- Brewster
- Chatham
- Chelmsford
- Clarksburg
- Cohasset
- Dunstable
- Duxbury
- Eastham
- Erving
- Groton
- Hadley
- Hamilton
- Hanover
- Harvard
- Hingham
- Holden
- Holliston
- Hopkinton
- Lenox
- Leverett
- Lincoln
- Longmeadow
- Lunenburg
- Lynnfield
- Manchester
- Marblehead
- Medway
- Mendon
- Middleton
- Natick
- Needham
- Newton
- Northborough
- North Reading
- Norwell
- Orleans
- Paxton
- Pelham
- Princeton
- Reading
- Richmond
- Rutland
- Sandwich
- Scituate
- Sharon
- Shrewsbury
- Shutesbury
- Southborough
- Sterling
- Swampscott
- Topsfield
- Upton
- Walpole
- Wayland
- Wellfleet
- Wenham
- Westborough
- West Boylston
- Westford
- Westwood
- Wilbraham
- Winchester

Highest 5 percent
(scores 2960–3075)

- Acton
- Andover
- Bolton
- Boxborough
- Brookline
- Carlisle
- Concord
- Dover
- Lexington
- Medfield
- Sherborn
- Stow
- Sudbury
- Wellesley
- Weston

[a]Average test score is sum of eighth-grade math and reading scores on the MEAP tests, averaged over the even-numbered years from 1988 through 1994 (missing years omitted, average adjusted for statewide trend). N = 321; 30 communities with total school enrollment less than 150 in 1980 are not included. Members of regional school districts (for eighth grade) are assigned their regional district's average eighth-grade score.

Appendix Table A2
Variables Used in the Analysis
(Number of observations = 321 Massachusetts cities and towns)

Variables:	Mean	Standard Deviation	Minimum	Maximum
Percent change in enrollment, grades 1–8				
1980–85	−17.1	10.5	−50.8	21.6
1990-95	14.2	12.3	−40.0	62.3
Predicted percent change in enrollment, grades 1–8[a]				
1980–85	−17.3	10.9	−42.9	15.0
1990–95	6.9	10.4	−19.8	41.4
Net change in enrollment, grades 1–8 (percentage points)[b]				
1980–85	.2	10.3	−32.2	45.7
1990–95	7.2	14.1	−77.8	68.0
School quality index[c]	2.71	.14	2.22	3.07
Dummy Variables: Prop 2-1/2 revenue cuts[d]				
for one year	.38	.49	0	1.00
for two or three years	.10	.30	0	1.00
Dummy variable: At levy limit in fiscal year 1989[e]	.40	.49	0	1.00
Developable land per housing unit[f]	.96	.93	.04	9.31
Median family income ($000)				
1980	20.0	5.3	10.1	47.6
1990	42.3	12.0	20.5	95.1
Percent of resident adults college educated				
1980	19.9	11.0	4.9	59.6
1990	26.7	12.7	6.9	65.5
Dummy Variable: Located inside Boston metropolitan area	.33	.47	0	1.00
Median house value ($000)				
1980	51.8	17.1	29.0	143.6
1990	169.7	57.4	88.9	497.9
Median unit rent ($)				
1980	272	55	116	501
1990	597	122	248	1001

Notes:

a. Predicted change in enrollment is the ratio of (number of MA residents who were age 1 to 8 at beginning of period) to (number who were age 6 to 13 at beginning of period) minus 1, expressed as percent.

b. Net change is actual percent change in enrollment minus predicted change, in percentage points.

c. School quality index is average student's combined math and reading test score averaged over the years 1988 through 1994, expressed in thousands.

d. Initial property tax cuts were made in fiscal years 1982 (one year), 1983, and 1984.

e. Communities are "at levy limit" if levy is greater than or equal to 99.9 percent of levy limit.

f. Developable land is acres of residential and nonpublic open land in 1984 divided by number of single family housing units in 1980.

Source: Massachusetts Department of Revenue; U.S. Bureau of the Census; Massachusetts Department of Education.

Globalization and U.S. Inflation

Geoffrey M. B. Tootell

Vice President and Economist, Federal Reserve Bank of Boston. *The author wishes to thank Lynn Browne and Dick Kopcke for their useful comments and suggestions. Valuable research assistance was provided by Rokeya Kahn and Delia Sawhney.*

Where has all the inflation gone for what is getting to be a long time passing? Estimates of the Phillips curve suggest that the low level of unemployment over the last few years should have produced a fairly significant acceleration in prices, yet inflation has continued to decline. Some, like Robert Gordon (1997) and Staiger, Stock, and Watson (1997), take this occurrence as evidence that the non-accelerating-inflation rate of unemployment, the NAIRU, has declined. Others argue that special factors, such as recent movements of employee health coverage to health maintenance organizations, have temporarily masked the increase in inflation. Yet, perhaps the most widely cited explanation for the surprisingly good inflation performance of late concerns the increasing sensitivity of the U.S. economy to foreign economic conditions; specifically, since capacity utilization abroad has been slack in recent years, U.S. inflation has remained mild. This study uses a variety of approaches to examine whether U.S. inflation depends on foreign, rather than domestic, capacity constraints. It is shown that foreign capacity plays little, if any, role in the determination of U.S. inflation independent of any role it might play in U.S. capacity utilization.

Monetary policymakers must understand the determinants of inflation in order to attain their inflation goal. If foreign capacity constraints help determine domestic price increases, then U.S. policymakers should modify their concerns about domestic capacity utilization. Going beyond, or falling below, some domestic measure of full capacity would not be sufficient to increase, or decrease, inflation, and the concept of a domestic NAIRU might not necessarily be an important or even meaningful policy goal. The rationale for the importance of foreign capacity to U.S. inflation is fairly clear-cut. If, for example, domestic demand exceeded domestic capacity, while foreign capacity remained underutilized, either the excess domestic demand for goods would be absorbed by imports or profits would be squeezed as labor markets tighten and costs rise. Little pressure on prices would result, as attempts by domestic producers to raise prices would immediately decrease demand for domestic goods and relieve pressure on costs; all this without monetary policy action. Similarly, if the U.S. economy were below full employment, while the rest of the world was beyond it, U.S. inflation would tend to rise.[1] In such a world, monetary policymakers would be less likely to achieve both their inflation and their output targets, since they would no longer control a key determinant of U.S. inflation.

Estimates of the Phillips curve suggest that the low level of unemployment over the last few years should have produced a fairly significant increase in the rate of inflation, yet inflation has continued to decline.

This article begins with a simple description of how global capacity could affect domestic inflation. Several measures of foreign excess capacity are then constructed in order to test the relationship empirically. As "excess capacity" is not directly measurable, several proxies are calculated. First, Phillips curves are estimated for our major trading partners and their NAIRUs calculated, in order to produce gauges of unemployment rate gaps. Additionally, deviations from trend GDP in those countries are examined. Neither of these measures of excess foreign capacity was found to have a direct effect on U.S. inflation.

Concerns about the precision of these estimates of foreign capacity utilization motivate another approach. Several empirical relationships should hold if foreign capacity affects U.S. inflation: Specifically, foreign capacity should affect U.S. import prices, and U.S. import prices should affect U.S. inflation. There is little evidence in support of either of these two links. In total, the results indicate that the Phillips curve, relating some measure of U.S. capacity utilization to U.S. inflation, is alive, if ailing a bit, even as the world gets more integrated.

I. Channels Through Which Foreign Capacity Might Affect U.S. Inflation

The posited link between global capacity utilization and U.S. inflation is straightforward. Foreign capacity utilization is assumed to affect the prices of foreign goods. Foreign goods prices, then, help determine U.S. import prices. And, finally, U.S. import prices may affect U.S. inflation through a variety of channels, both direct and indirect.

The direct effect of import prices on U.S. inflation is simple arithmetic. The inflation measure examined throughout this article is changes in the Consumer Price Index less food and energy (core CPI) as it captures the goods most directly related to the public's welfare.[2] Many foreign goods prices are included in the CPI. Thus, CPI inflation,

$$\pi_i^{CPI} = \alpha * \pi_i^M + (1 - \alpha) * \pi_i^{NM}, \qquad (1)$$

is a weighted average of inflation in domestically produced goods prices, π^{NM}, and inflation in imported goods prices, π^M, where α is the share of imported goods in the CPI. Equation 1 illustrates that any change in the dollar price inflation of foreign goods at the consumer level will be directly captured by the domestic inflation measure. Note that this direct effect makes the CPI more likely than the GDP deflator to exhibit any effects from foreign capacity utilization.[3]

The inflation of foreign goods prices could also have several indirect effects on U.S. goods price inflation. First, and perhaps most important, price movements of foreign goods should affect the prices of U.S. goods with which they directly compete. In fact, much of the recent debate attributes the surprisingly low current inflation to foreign competition. It has been argued that U.S. producers cannot raise their prices even when cost pressures begin to appear, because doing so when foreign prices remain moderate would seriously diminish their market share. Thus, the smaller the U.S. producers are in this traded goods market, or the more substitutable the goods, the greater the discipline imposed by foreign prices. This greater discipline could manifest itself in lower wage

[1] The response of the exchange rate to these different capacity utilization rates is usually left rather unclear in these arguments. Exchange rate movements should moderate any effects on domestic inflation. In fact, inflation rates routinely vary across countries, as do changes in these rates, with exchange rate movements helping to offset these differences.

[2] Most of the conclusions in this paper do not depend on which price measure is used, however. The exceptions will be discussed as they arise.

[3] The GDP deflator does not include imports directly. The effect of import prices on both U.S. input costs and U.S. export prices, which are included in the deflator, might produce some reaction of the deflator to foreign capacity utilization. However, this effect would be much less direct than the effect on CPI inflation.

demands by workers in these domestic firms, in a profit squeeze, or in a decline in U.S. production as a response to rising imports; any of these alternatives would offset or relieve the pressure of U.S. capacity constraints on U.S. prices. Second, movements in foreign goods prices can indirectly affect U.S. goods prices if those foreign goods are important inputs to U.S. goods production. Fluctuations in the prices of foreign goods used as inputs, and particularly permanent changes in these prices, alter the costs, and thus the prices, of U.S.-produced goods and services.

Foreign capacity plays little, if any, role in the determination of U.S. inflation independent of any role it might play in U.S. capacity utilization.

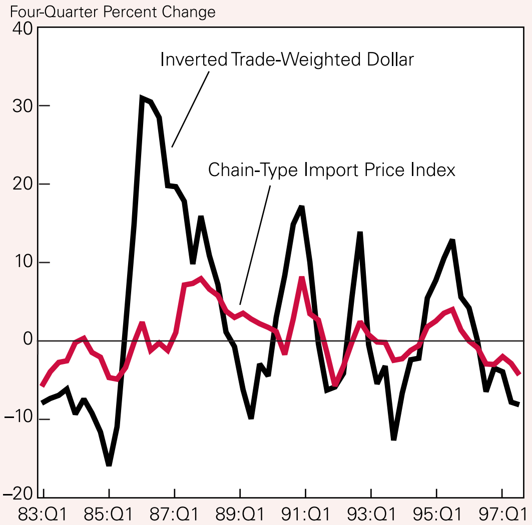
Without these two major indirect channels, global capacity constraints would affect only a small subcomponent of U.S. inflation, since imports represent a small share of total U.S. output. If only the direct effect were operative, inflation for U.S. produced goods and services would still essentially be determined by domestic capacity constraints. Thus, a finding that import price inflation affects the U.S. inflation rate need not imply that foreign goods are acting as a constraint on the price inflation of U.S.-produced goods or that foreign capacity determines the rate of U.S. inflation. For foreign capacity utilization to be a significant determinant of U.S. inflation, import prices must affect the price inflation of domestically produced goods, π^{NM}.

Nevertheless, for several reasons, one might expect foreign capacity utilization to have little effect on U.S. inflation. Movements in the dollar value of the production costs faced by foreign producers, resulting from a change either in their home currency costs or in the value of the dollar, may have little effect on the prices of imported goods in the United States. Several margins—the manufacturer's, the importer's, the retailer's—may be squeezed or expanded by such price changes. For a variety of reasons foreign producers may let profits bear the brunt of any change in the dollar value of production costs. If, for example, the dollar appreciates—reducing the dollar value of foreign costs—foreign firms might allow profits to increase rather than cut U.S. prices, either because they are price followers in the U.S. market or because they perceive the appreciation as temporary. Changes in costs resulting from excess capacity abroad may also be offset by movements in the dollar. Finally, even if changes in foreign prices or the value of the dollar did affect import prices, this change in import prices may have little indirect effect on U.S. goods prices. U.S. and foreign goods may be imperfect substitutes, so that domestic firms do not find that foreign prices greatly affect the demand for their products.

In fact, there is abundant evidence that changes in exchange rates, and thus changes in the dollar costs faced by foreign producers, have only a modest effect on the dollar prices foreign producers charge in the United States. The debate over the degree of "pass-through" has received considerable attention since the mid 1980s when, as shown in Figure 1, dollar import prices failed to fluctuate nearly as much as the value of the dollar itself. The experiences in the mid 1980s resulted in several theoretical explanations for only a

Figure 1

Trade-Weighted Dollar and Import Prices



Source: Trade-weighted dollar data, from G.5 release, Federal Reserve Board. Chain-type import price index, from U.S. Bureau of Labor Statistics.

partial pass-through.[4] This literature emphasized that large foreign players here, firms that face a downward-sloping demand curve, will pass through only part of a change in the dollar value of their production costs to their U.S. prices. Of course, if foreign producers are small players in the U.S. market, they will tend to take dollar prices as given, with very little pass-through. This theoretical work has also received broad empirical support at the micro level.[5] The industry-level data suggest that foreign producers are likely to change their U.S. prices significantly only if their U.S. competitors do so also, and U.S. competitors alter their prices when strains in capacity in the United States raise their costs. These results imply that import prices should be correlated with capacity utilization in the United States, not capacity constraints abroad.

Much of the recent debate attributes the surprisingly low current rate of inflation to foreign competition.

Since theory is agnostic about whether global capacity determines U.S. inflation, it is left to empirical analysis to answer the question. Accordingly, the next few sections examine the data. Foreign capacity since the early 1970s is calculated, first by estimating the NAIRU in each of our six major trading partners and, alternatively, by taking a simple trend of real GDP in these countries. It is then shown that deviations from full employment or trend GDP in these trading partners have had no apparent effect on U.S. inflation. Examining the various possible sources of the break in the transmission mechanism could help sidestep the problems with measuring foreign capacity. Either foreign capacity could have little effect on foreign prices, foreign prices could have little correlation with U.S. import prices, or U.S. import prices could have no effect on U.S. inflation. The Phillips curve regressions for the foreign countries in the next section show that the link between foreign capacity utilization and foreign inflation is fairly strong. The subsequent section reveals that the link is weak between foreign inflation and inflation in U.S. import prices. But perhaps more important, U.S. import price inflation fails to have a significant effect on U.S. CPI inflation, and this last result is not dependent on any particular measure of foreign capacity utilization.

II. Measuring Foreign Capacity

In theory, measuring excess capacity is relatively straightforward. Full capacity is defined as the level of GDP or unemployment at which all of the economy's available resources are utilized voluntarily at stable wages and prices (or stable rates of wage and price inflation). Calculating the output level when all inputs are so utilized, and comparing that level to the actual output level of the economy at any given time, is one way of discerning whether the economy is running at, below, or above full capacity. However, actually ascertaining the amounts of labor and capital available to the economy at full employment and determining their productivities is problematic. Since it is almost impossible to measure capacity in this way, other approaches are usually taken.

One commonly used alternative calculates potential output as an average of past actual values, once allowances are made for stable growth. It is argued that, over the long run, prices and wages adjust to ensure that the economy is at full employment. The economy may occasionally operate above or below its capacity, but on average it should run near its potential level. In reality, output may differ from its full-employment level for a significant length of time, so this method of calculating potential output could produce a serious over- or underestimation of capacity.[6] Shifts in potential output, or output growth, also confound this approach; including output levels from the period before any such shift will bias the estimate of potential away from its true value. These failings aside, this is one approach used in this study.

Inflation is an important indicator of whether the economy is above or below full capacity, one that the previous approach ignores. If the economy is operating beyond (below) capacity, inflation will tend to rise (fall). Estimating a country's level of full employment by estimating its Phillips curve captures the informa-

[4] Examples of this theoretical discussion can be found in Baldwin (1988) and Dixit (1989).

[5] Empirical studies of pass-through can be found in Feenstra, Gagnon, and Knetter (1996), Gagnon and Knetter (1995), and Froot and Klemperer (1989), for example.

[6] Many countries in Europe have operated well below full employment through most of the 1980s and 1990s. As a result, such estimates of potential output in these countries will tend to be biased down.

Table 1
The Phillips Curve: Six Major U.S. Trading Partners

	Canada	France	Germany	Italy	Japan	United Kingdom
Constant	1.11*	.41*	.18	1.91*	.39	1.41*
Unemployment	−.13*	−.23	−.04*	−.85*	−.47*	−.71*
Lagged Inflation	1.0	1.0	1.0	1.0	1.0	1.0
Estimated NAIRU	8.50*	6.91*	5.10*	9.50*	2.30*	7.30*
	(.35)	(1.33)	(1.17)	(.77)	(.50)	(.93)
Log Likelihood	−68.0	−56.4	−55.6	−117.2	−133.3	−151.3

Note: The coefficients on 2 lags of the unemployment rate are summed. The sum of coefficients on 12 lags of inflation is constrained to 1. The sample period is from 1971 to 1996.
Standard errors are in parentheses.
*Significant at the 5 percent level.
Source: Inflation and unemployment data for the six countries were obtained from the OECD, *Main Economic Indicators*.

tion about capacity utilization contained in the inflation numbers.[7] Defining full capacity by its effect on inflation is also the relevant measure for the focus of this study, since foreign capacity affects U.S. inflation through its effect on foreign inflation. For this reason, estimates of foreign capacity derived from estimates of the Phillips curve for each of our six major trading partners are also used to examine the effect of foreign capacity utilization on U.S. inflation. As a by-product, estimating foreign Phillips curves tests whether foreign capacity affects foreign inflation, which is the first link in the chain between foreign capacity constraints and U.S. inflation.

Data limitations on foreign prices and unemployment rates make it impossible to estimate this relationship for all of our trading partners.[8] The most reliable data belong to our six major developed trading partners, Canada, France, Germany, Italy, Japan, and the United Kingdom, so capacity utilization in these countries is examined. Although these countries consistently account for roughly 50 percent of all U.S. imports throughout the sample period, it is likely that their capacity utilization would have a disproportionately important effect on U.S. inflation, since imports from these developed economies are more likely to compete directly with U.S. goods production. Thus, these imports are more likely to have both a direct and an indirect effect on U.S inflation.

The estimates of a simple quarterly Phillips curve for each of our six major trading partners from 1971 to 1996 are presented in Table 1. The basic specification of the Phillips curve equation,

$$\pi_{tj} = \alpha + \Sigma\beta_j * UR_{t-i,j} + \Sigma\gamma_{t-i,j} * \pi_{t-i,j} \quad (2)$$

for j = *Canada, France, Germany, Italy, Japan, U.K.*

was allowed to vary slightly for the six different countries, although imposing the same specification across countries had no effect on the results. In all cases the γ coefficients on the lagged domestic inflation realizations were constrained to sum to one. However, a nonlinear relationship between the unemployment rate and inflation was assumed for Italy, Japan, the United Kingdom, and France; in these countries a log linear specification appeared to fit the data better. Further, the speed of the response of inflation to the unemployment rate was allowed to vary by country; the number of quarterly lags of the unemployment rate ranged from two to three.

The second row of Table 1 reveals that each country's unemployment rate had a strong negative effect on that country's rate of inflation. In all but France, the unemployment rate was a significant determinant of foreign inflation, and even in France the unemployment rate was marginally significant.[9]

[7] One drawback to a simple Phillips curve is that other forces besides each country's domestic capacity could affect inflation, such as oil price shocks or the sources investigated in this paper, import prices.

[8] Data on unemployment and inflation rates for some of the currently important exporters to the United States, such as Mexico and China, are unreliable or unavailable for much of the sample. These countries were also much smaller exporters to the United States in the early part of the 1971–96 sample.

[9] The data show that with more than 90 percent certainty the unemployment rate in France is an important determinant of its domestic inflation.

The next to the last row of Table 1 presents the estimated NAIRUs for each of these countries. They range from a low of 2.3 percent in Japan to a high of 9.5percent in Italy. These estimates seem to coincide with conventional wisdom; Japan's unemployment rate rarely rose above 3 percent, and the relatively high estimated NAIRUs for France, Italy, Germany, and the United Kingdom are consistent with the diagnosis of Eurosclerosis in the 1980s. The estimate of Canada's NAIRU at 8.5 percent also seems reasonable, as the high unemployment rate that country has experienced over the last several years has reaped a relatively slow payoff in lower inflation.

More important, even for such parsimonious Phillips curves, the NAIRUs are estimated with relative precision.[10] This fact is important since the more precise the estimate of the NAIRU, the more reliable the estimate of foreign excess capacity. Based on the literature hypothesizing hysteresis of the NAIRU in Europe, such as Blanchard and Summers (1986) and Sachs (1986), one might expect the standard errors of the NAIRU estimates for the European countries to be quite large, given the shocks Europe experienced over the sample period. And, in fact, the estimates with the least certainty, France and Germany, are the two countries for which hysteresis in the NAIRU was most seriously discussed in the 1980s. Yet even the NAIRUs for these two countries are estimated fairly reliably. Furthermore, the estimates for Canada and Japan are the most precise, which is not surprising since little has been said about unstable NAIRUs in these countries.

The NAIRUs for our major trading partners are estimated with relative precision—important since the more precise the estimate of the NAIRU, the more reliable the estimate of foreign excess capacity.

Deviations of GDP from estimates of its potential level are used as an alternative measure of excess capacity in these countries. Potential output is calculated simply as the trend GDP over the previous 26 years. The results attained using deviations from potential GDP are robust to the inclusion of different trends over various parts of the sample, although the more parsimonious model is used in the estimation presented in the tables.

III. Foreign Capacity and U.S. Inflation

A standard Phillips curve for the United States,

$$\pi_t^{US} = \alpha + \Sigma\beta_i * UR_{t-i} + \Sigma\gamma_i * \pi_{t-i}^{US} + \Sigma\Theta_i * \pi_{t-i}^{NOM} + \Sigma\eta_i * \pi_{t-i}^{OIL} + \lambda * Nixon + \tau * Nixoff, \qquad (3)$$

variants of which are found in Gordon (1977, 1982, 1997), Motley (1990), Weiner (1993), Tootell (1994), and Fuhrer (1995), is used here to examine the effects of foreign capacity constraints on U.S. inflation. Foreign capacity utilization, $UR_t^* - UR^{*FE}$, should help determine non-oil import prices, π^{NOM},

$$\pi_t^{NOM} = \Sigma\beta_i^F * (UR_{t-i}^* - UR^{*FE}), \qquad (4)$$

which could affect the U.S. rate of inflation both directly and indirectly.[11] Plugging equation 4 into equation 3,

$$\pi_t^{US} = \alpha + \Sigma\beta_i * UR_{t-i} + \Sigma\gamma_i * \pi_{t-i}^{US} + \Sigma\varphi_i * (UR_{t-i}^* - UR^{*FE}) + \Sigma\eta_i * \pi_{t-i}^{OIL} + \lambda * Nixon + \tau * Nixoff, \qquad (5)$$

produces the base specification examined in Table 2.[12] As is usual, the coefficients on the lagged U.S. inflation variable are assumed to sum to one, to ensure inflation stability at a natural rate.[13]

The NAIRUs used to calculate the foreign excess capacity in equation (5) were derived from the results in Table 1. The measure of foreign excess capacity also

[10] In fact, examining different specifications for the Phillips curve, such as estimates including oil price shocks, did not significantly alter the estimates of the NAIRUs, even though their inclusion often helped lower the standard errors.

[11] Non-oil import prices could also depend on lagged prices and should depend on the exchange rate. Which prices to use—lagged import prices or lagged trade-weighted foreign prices, for example—is unclear. Both of these price measures were also included in this standard specification, along with the trade-weighted value of the dollar, with no effect on the results.

[12] Note that any effect of import prices on U.S. capacity will be captured in the estimation through the unemployment rate. Also, Nixon and Nixoff are dummy variables capturing the quarters when wage and price controls were instituted and released.

[13] A specification that examined relative price shocks, both import price inflation and oil import price inflation relative to U.S. inflation, was also tested. The results were identical.

Table 2
Foreign Capacity Utilization and U.S. Inflation

	Full Sample: 1973:III–1996:II			Short Sample: 1984:III–1996:II		
	Domestic Capacity	Foreign Capacity Added	Nonlinear Relation	Domestic Capacity	Foreign Capacity Added	Nonlinear Relation
Constant	.72*	.64*	.59	.40*	.28	.28
U.S. Unemployment	−.11*	−.10*	−.09*	−.06*	−.05*	−.05*
Lagged Inflation	1	1	1	1	1	1
Oil Import Price Inflation	.01*	.01*	.01*	.0001	.0003	.0001
Foreign Output Gap		−.05	−.30		−.08	−.56
Nixoff	.64*	.63*	.62*			
Observations	92	92	92	48	48	48
Log Likelihood	−17.88	−16.46	−15.96	42.76	45.89	46.4

Note: The coefficients on two lags of unemployment rate are summed. The sum of the coefficients on 12 lags of inflation is constrained to 1. Foreign Output Gap refers to the trade-weighted deviation of unemployment from the NAIRU in the six countries.
*Significant at the 5 percent level.
Source: Data for U.S. inflation, unemployment, and import prices are from the *Current Population Survey*, U.S. Bureau of Labor Statistics.

requires aggregation of the six different excess capacities. The weights chosen to aggregate these excess capacities were the shares of U.S. imports from each country.[14]

These weights should be a good indicator of the pressure on U.S. import prices and U.S. inflation from a given country's capacity, since that pressure should depend on the country's relative importance in U.S. imports. Both U.S. and foreign capacity are included in the equation. If foreign capacity alone determined U.S. inflation, then measures of U.S. capacity utilization should be insignificant and measures of foreign capacity significant. If foreign excess capacity merely moderates U.S. inflation, particularly since U.S. capacity is an important part of global capacity, then both measures of capacity utilization should be important.

Finally, the effect of oil price movements on U.S. inflation is accounted for separately in these Phillips curve estimations, even though the equations examine the CPI excluding food and energy. The inflation in oil prices must be included because oil is an important input to production of so many goods; it should be included separately from other import prices because, to a large extent, changes in oil prices over this sample period were the result of oligopolistic price-setting behavior that was primarily a function of global politics, not rates of global capacity utilization. As a result, the effect of foreign capacity utilization on U.S. inflation would occur mostly through non-oil import prices.

The coefficients from the estimation of variants of equation (5) are presented in Table 2. The equation is estimated from 1973 to 1996 because the country-specific import weights required to calculate the foreign capacity utilization rate only go back that far.[15] The base specification includes 2 quarterly lags of the unemployment rate, 12 quarterly lags of the inflation rate, and 4 quarterly lags of oil import price inflation. The first column contains the coefficients from a standard Phillips curve—one estimated without foreign capacity. As expected, U.S. unemployment has a significant negative effect on U.S. inflation, and oil prices affect inflation in the expected direction. Since the sample period includes the end of the Nixon wage and price controls, an indicator variable capturing the effect of removing these controls is also included in the regression; its coefficient is correctly signed and of marginal significance.

The second column of Table 2 presents the coefficients from a Phillips curve estimation that adds two lags of the trade-weighted excess of each country's

[14] The shares were based on total U.S. imports from these six countries, so they always summed to one. As a result, there is no trend in the foreign excess capacity variable due to any possible trend in the weights, although this share appeared to be relatively constant around 50 percent.

[15] A longer sample was used when estimating the foreign NAIRUs since the import weights were not needed. The longer sample provides more information about each country's NAIRU.

unemployment above its estimated NAIRU. If foreign capacity constraints significantly determine U.S. inflation, either directly or indirectly, the coefficient should be negative and significant. It is negative but far from significant.[16] Concerns about collinearity between the U.S. and foreign capacity measures can be assuaged, as the estimates of the other coefficients in the equation are little affected by the inclusion of the foreign unemployment gap. In fact, the estimated effect of U.S. unemployment on U.S. inflation remains essentially unchanged from column 1, as is its level of significance. It is not that we cannot tell the difference between the domestic and foreign capacity measures, but that the foreign measure does not appear to affect U.S. inflation.

The exact functional form of any possible relationship between foreign capacity and U.S. prices is not clear. For this reason, column 3 reestimates the Phillips curve allowing for a nonlinear connection between U.S. inflation and foreign capacity. Estimating a nonlinear specification does not increase the importance of foreign capacity. The size of the coefficient is larger in absolute terms because of the change in the functional form.[17] The insignificance of foreign capacity is unaffected by this change in the specification.[18] Other functional forms were tested, with the same result. There is little compelling evidence that this measure of foreign capacity utilization affects either U.S. inflation or the relationship between U.S. capacity and U.S. inflation.

It has been asserted that the importance of foreign capacity to U.S. inflation increased with the trade gap in the 1980s. If so, foreign capacity utilization may have become significantly more important in the mid 1980s. To ensure that the previous regressions are not masking the effect of foreign capacity on U.S. inflation by including the 1970s, the final three columns of Table 2 present the coefficients from reestimates of the first three columns over a shorter sample period, from 1984 on.[19] Foreign capacity remains insignificant.[20] The inclusion of the foreign capacity variable still has

It appears that foreign capacity utilization has no effect on U.S. inflation, and that the relationship has not strengthened since the 1980s.

little effect on the other coefficient estimates. It appears that foreign capacity utilization has no effect on U.S. inflation, and that the relationship has not strengthened since the 1980s.

As mentioned earlier, one's confidence in the measure of foreign excess capacity in Table 2 depends on the reliability of the estimates of the foreign NAIRUs. To examine the robustness of the results to different measures of foreign capacity utilization, Table 3 repeats Table 2 using the deviation of GDP from its trend level for both the United States and, on a trade-weighted basis, the six foreign countries examined in this study. The same six countries are used in order to maintain consistency with the previous table.

The results are identical. The first column of Table 3 shows that estimates of a simple Phillips curve using the U.S. GDP gap are very similar to one using the unemployment rate. The positive and significant coefficient on the U.S. GDP gap indicates that GDP beyond trend significantly increases U.S. inflation, as expected. Oil prices remain significant in this reformulation of the basic Phillips curve. Including the trade-weighted deviation of foreign GDP from its trend,

[16] The significance is unaffected by the number of lags of foreign capacity included in the equation. Since two lags of the trade-weighted foreign capacity variable are included, the log likelihood ratio testing whether we can reject that the coefficients on the foreign capacity are equal to zero is distributed as a chi-square with two degrees of freedom. The critical value for this ratio is 7.38. Its actual value is 2.8, providing little support that the coefficients are different from zero. In general, this result holds for other inflation measures. For the preferred specification of the chain GDP deflator, the total CPI, core PPI, and the deflator on personal consumption expenditures, foreign capacity has no statistically significant effect. Only for total PPI does foreign capacity appear significant. However, there is little evidence of a robust relationship between these foreign capacity measures and U.S. inflation.

[17] In column 3, U.S. inflation is estimated as a linear function of the U.S. unemployment rate and a nonlinear function of the trade-weighted foreign unemployment gap; as a result, the coefficients are not comparable. The linear form is maintained for the U.S. unemployment rate, as it appears to fit the data better.

[18] The critical value for the significance of the coefficient on foreign capacity is 7.38, and the log likelihood ratio remains far below that value, at 3.8.

[19] The NAIRUs of our six major trading partners were reestimated over the shorter sample because of the concerns that the NAIRUs in Europe rose substantially in the shorter period. Only the estimate of the NAIRU in Italy rose significantly. The results are identical when the NAIRU estimates from the longer sample are used.

[20] The critical value for accepting the importance of foreign capacity utilization in U.S. inflation in the shorter sample is 7.38. The actual value of the log likelihood ratio is 6.26, rejecting that foreign capacity plays an important role in the determination of U.S. inflation. The rejection is stronger if the sample begins in 1980.

Table 3
The U.S. Phillips Curve, with Trade-Weighted Output Gaps

	Full Sample 1973:III–1996:II		Short Sample 1984:III–1996:II	
	Domestic Capacity	Foreign Capacity Added	Domestic Capacity	Foreign Capacity Added
Constant	−.07*	−.07*	−.08	−.05
U.S. Output Gap	9.17*	8.07*	7.25*	4.1*
Lagged Inflation	1.0	1.0	1.0	1.0
Oil Import Price Inflation	.01*	.01*	0	0
Foreign Output Gap		−0.21		1.30
Nixoff	.90*	.92*		
Log Likelihood	−20.97	−18.93	−16.23	−14.75

Note: Nixoff is a price control for the period 1974:Q2–1975:Q1. U.S. Output GAP refers to the deviation of actual income from its potential, and foreign output gap is the trade-weighted deviation of actual income from potential income in the six countries.
*Significant at the 5 percent level.

Source: U.S. GDP data from the U.S. Bureau of Economic Analysis. Foreign GDP data from the Bank of International Settlements, Switzerland.

column 2, has little effect on the estimates in column 1; the coefficient estimate on the U.S. GDP gap is roughly the same, and the coefficient on foreign capacity utilization is small, insignificant, and incorrectly signed.[21] The last two columns of Table 3 show that the results, again, do not change if the shorter sample is used.[22] There appears to be little evidence that foreign excess capacity, no matter how it is measured, helps determine U.S. inflation.

IV. The Effect of Foreign Capacity on Import Prices

It is still possible that the results reveal more about these two measures of foreign capacity than about the effect of foreign capacity on U.S. inflation. For this reason, a different approach is also taken, one that attempts to sidestep the difficulties of quantifying foreign excess capacity. This section and the next examine the different stages of the transmission mechanism through which foreign capacity should affect U.S. inflation. The first stage of the transmission of foreign capacity to U.S. inflation is the effect of foreign capacity on inflation in foreign prices. The foreign Phillips curves in Table 1 show that foreign capacity utilization does affect foreign rates of inflation. Thus, the progression from foreign capacity utilization to U.S. inflation is being broken either at the link between foreign inflation and U.S. import prices or at the link between U.S. import prices and U.S. inflation. This section examines the first of the latter two links.

Attempts to explain U.S. import prices are presented in Table 4. These regressions are less grounded in theory or history than the Phillips curve equation, so the results from several specifications were examined. The basic formulation of U.S. non-oil import price inflation,

$$\pi_t^{NOM} = \alpha + \Sigma\beta_i{*}GAPUS_{t-i} + \Sigma\Theta_i{*}GAPF_{t-i} + \Sigma\gamma_i{*}\pi_{t-i}^{NOM} + \Sigma\eta_i{*}\pi_{t-i}^{OIL}, \qquad (6)$$

includes the U.S. and the trade-weighted foreign capacity gaps, GAPUS and GAPF, measured either as deviations from trend GDP or as deviations from the estimated NAIRU, and π^{NOM}, non-oil U.S. import price inflation. Thus, non-oil import price inflation is explained using lags of itself, oil import price inflation, the trade-weighted foreign GDP or unemployment gaps, and the U.S. GDP gap or U.S. unemployment rate.[23]

The justification for including the foreign gap is clear; if foreign excess capacity is holding down foreign costs of production, it might restrain the price inflation of foreign goods to U.S. consumers and producers. Alternatively, the level of U.S. capacity might affect import prices, since foreign goods may be

[21] The log likelihood ratio is distributed as a chi-square with two degrees of freedom. Its critical value is 7.38, while the ratio's actual value is 4.08.

[22] The log likelihood ratio is again distributed as a chi-square with two degrees of freedom. Its critical value remains 7.38, while its actual value over this shorter sample is 2.96.

[23] All the import price inflation regressions were also estimated including the trade-weighted exchange rate. These results are less relevant for this study since it is the total derivative of import prices and foreign capacity utilization that we are concerned about, not its partial derivative holding the change in the exchange rate constant. However, when the exchange rate was included in the import price inflation regressions reported in the text, the foreign capacity variable was more apt to reveal a significant effect on U.S. import prices, although the effect was not very robust to different samples or specifications.

Table 4
Domestic and Foreign Capacity Utilization and U.S. Import Price Inflation

	Full Sample 1974:I–1996:II		Short Sample 1984:I–1996:II	
	Unemployment	GDP GAP	Unemployment	GDP GAP
Constant	−.12	.29	−1.26	.18
U.S. Output Gap	.087	6.825	.21	3.929
Foreign Output Gap	−.29	−6.95	.078	−2.42
Non-Oil Import Price Inflation	.452*	.4849*	.5855*	.494*
Oil Import Price Inflation	.045*	.0552*	.0058	−.01004
Observations	90	90	50	50
Log Likelihood	−143.65	−146.06	−70.2	−65.01

Note: The coefficients on 2 lags of the unemployment rate are summed. The sum of the coefficients on 12 lags of inflation is constrained to 1.
*Significant at the 5 percent level.

priced to market. The lagged oil prices are examined for the same reason as in the U.S. Phillips curve analysis: Inflation in oil prices tends to affect the prices of all goods because oil is such a ubiquitous input to production. The lags of non-oil import price inflation capture the inertia in the inflation rate of these prices, but their inclusion is far less justifiable than that of the lags of the dependent variable in the usual Phillips curve equation, which are usually assumed to capture inflation expectations. As a result, the coefficients on these lagged prices, the γs, are not constrained to sum to 1, particularly since foreign prices and wages will be determined by the expectations of foreign prices in general, rather than simply the prices of their exports to the United States.

The first column in Table 4 shows that neither foreign nor U.S. capacity adds independently to the explanation of import prices when both are included. The significance tends to be higher for the foreign measure of excess capacity. The second column reveals that the same is true when excess capacity is measured as a GDP gap. The final two columns highlight the fact that this effect did not strengthen over the most recent period. Only the coefficients on the lagged inflation terms are significant in these regressions.

As mentioned above, it is unclear which lagged inflation rate should be included as an explanatory variable for import price inflation. One might argue that, just as in the Phillips curve estimation, foreign inflation in general should be included, since the determinant of the export prices foreigners charge U.S. customers will depend on the expectation of their own inflation rates; firms care about their real return relative to their domestic currency. In this case, lags of the trade-weighted inflation rate as a whole should be included as an explanatory variable. When lags of non-oil import price inflation are replaced with lags of the trade-weighted foreign price inflation, both U.S. capacity, when measured as a GDP gap, and foreign capacity, when measured as an unemployment gap, tend to have a significant effect on U.S. import prices. All in all, however, although there is some evidence that foreign capacity has an effect on U.S. import prices, it is not particularly robust to different specifications, samples, or the measure of capacity utilization used.

V. The Effect of Import Prices on U.S. Inflation

Examining the determinants of U.S. import price inflation sheds little light on the insignificance of foreign capacity for U.S. inflation. Thus, the last stage in the transmission of foreign capacity's effect on U.S. inflation is examined in this section. The importance of non-oil import price inflation to U.S. inflation can be directly estimated in the U.S. Phillips curve. If foreign capacity affects U.S. inflation, it must be the case that U.S. import price inflation influences U.S. inflation, either directly or both directly and indirectly. Furthermore, if import prices do significantly affect U.S. inflation, then the weakness of the link between foreign capacity utilization and U.S. import prices would appear to explain the failure of foreign capacity to affect U.S. inflation.

Table 5 presents the coefficients from a simple Phillips curve regression with and without various measures of import price inflation. The first column provides the estimated coefficients of the base specification when oil price inflation is included and non-oil import price inflation is omitted. The sample is extended back to the late 1960s, since the availability of

Table 5
Phillips Curve and Import Prices, 1968–1996

	Oil Import Prices	Oil & Non-Oil Import Prices
Constant	.528*	.458*
U.S. Unemployment	−.0876*	−.084*
Lagged Inflation	1.0	1.0
Oil Import Price Inflation	.0105*	.0086*
Non-Oil Import Price Inflation		.047
Nixon	−.30*	−.35*
Nixoff	.61	.52
Log Likelihood	−18.26	−15.0

*Significant at the 5 percent level.

country-specific import data is no longer a binding constraint.[24] The unemployment rate is significant and correctly signed, as is the sum of the coefficients on the oil import price inflation.

The second column of the table presents the coefficient estimates when both oil import price inflation and non-oil import price inflation are included.[25] Oil import price inflation still adds significantly to the explanation of U.S. inflation, while non-oil import price inflation does not.[26] These results suggest that any effect of import prices on U.S. inflation derives from their oil component. If, however, import prices were having a significant effect on U.S. inflation, either directly or indirectly, both oil and non-oil import prices should influence U.S. inflation. It appears that the only foreign price changes to significantly affect U.S. inflation were oil price changes.

The lack of a significant relationship between import price inflation and U.S. inflation is surprising.[27] Common sense, and equation 1, tell us that if import prices change radically enough, they must have some effect on U.S. inflation. Table 6 examines this hypothesis by distinguishing episodes when non-oil import price inflation changed significantly from those when the movements were more benign. Non-oil import price inflation is defined as significant if it is a standard deviation above or below its mean value. The first column shows that large changes in non-oil import prices do have a significant effect on U.S. inflation. Columns 2 and 3 examine whether this effect is symmetric: Do large decreases in import prices have the same effect as large increases in import prices? The last two columns of Table 6 reveal that all the significance of the threshold effects comes from large *increases* in non-oil import price inflation. This result is interesting because it is relevant for today's policy debate since import prices are declining, not increasing. There is no evidence that even large declines in import prices affect U.S. inflation.

Large increases in non-oil import price inflation do have a significant effect on U.S. inflation, just as foreign oil price changes do, but large declines in these prices do not.

As imports are a component of the CPI, one would expect that the coefficient on import prices

[24] The results are identical over the sample examined in Tables 2, 3, and 4, however.

[25] The constraint on the coefficients on the lagged price variables in the Phillips curve is invalid if a subcomponent of the index is run independently. However, as will be discussed later, the import deflator prices do differ from the consumer prices, so the import prices are not exactly a subcomponent. Furthermore, examining import price inflation relative to U.S. inflation has no effect on the results.

[26] Since four lags of all the import price inflation indexes were examined, the likelihood ratio is distributed as a chi-square with four degrees of freedom, the critical value of which is 11.1. The actual value for the log likelihood ratio for the test of the significance of the coefficients on oil import price inflation is 16.02. The value of the likelihood ratio testing whether non-oil import price inflation adds to the Phillips curve is 6.52, well below the critical value of 11.1.

[27] The tests in Table 5 were performed with other measures of U.S. inflation. The results examining total CPI and the deflator for consumer expenditures were consistent with the results above; a specification with a statistically selected lag length of past inflation rates always rejects the importance of non-oil import price inflation in these other consumer price measures. The importance of the non-oil import prices in the PPI is consistently rejected over various specifications and samples. Over some specifications, there is some evidence that non-oil import prices influenced inflation of the GDP deflators, both the fixed-weight and the chain-weighted deflators. Since GDP inflation measures include no imported goods directly, the result might appear surprising. However, the close correlation between export and import prices could explain the result. Even with the GDP deflators, the importance of non-oil import prices is not robust to different specifications of the lags in the prices used or the sample selected. Other inflation measures do not offer strong evidence of a relationship between non-oil import prices and U.S. inflation.

Table 6
Phillips Curve and Import Prices: Threshold Values

	Large Changes in Non-Oil Import Prices	Large Movements Up in Non-Oil Import Prices	Large Movements Down in Non-Oil Import Prices
Constant	.498*	.57*	.519*
Unemployment	−.087*	−.1011*	−.085*
Lagged Inflation	1	1	1
Oil Import Price Inflation	.0065*	.0065*	.0114*
Large Changes in Non-Oil Import Prices	.0796*		
Large Upward Changes		.0967*	
Large Downward Changes			.0436
Nixon	−.32*	−.33*	−.31*
Nixoff	.397*	.305	.593*
Log Likelihood	−11.28	−11.2	−15.3

*Significant at the 5 percent level.

would at least be equal to their share in the CPI, while coefficient estimates greater than that share would suggest pricing discipline on U.S. producers. Although imports are a relatively small share of the CPI, if those prices consistently moved with import prices, the coefficient may be small, but the effect would be statistically significant. A closer examination of the two price series may explain why it is not. The CPI measures the prices paid by consumers at the retail level. The import price index measures the prices paid by the importer at the docks. The importer charges a different price to the distributor, the distributor charges another price to the retailer, and finally the retailer charges still another price to the consumer. The profit squeeze or profit expansion caused by the differential between import prices and U.S. consumer prices may be borne by any of these links in the chain. In other words, the pricing to market is not relevant just to the foreign producer's behavior, but it may also be important to the behavior of the domestic importers, distributors, and retailers. For this reason, the behavior of the dock price of imports may differ significantly from their retail price, especially in the short run.

VI. Conclusion

It has been asserted repeatedly that the surprisingly good news on inflation, given the low level of the unemployment rate, is not only good luck but the direct result of increased global competition and excess global capacity. If this statement is true, then policymakers need not worry about the current low level of unemployment since, given the current level of excess global capacity, inflation will not rise. However, if this assertion is correct, several relationships should be clear in the data. Most obviously, global capacity should directly affect U.S. inflation in a traditional Phillips curve. This paper finds no evidence of such a relationship. Furthermore, if foreign capacity did affect U.S. inflation, U.S. import prices would depend on foreign capacity, and U.S. import prices would have a strong relationship to U.S. consumer prices. Neither relationship appears to occur. The results in this study suggest that anyone who believes in a world where we no longer need worry about domestic capacity constraints will eventually be rudely awakened by data that suggest otherwise.

References

Baldwin, Richard. 1988. "Hysteresis in Import Prices: The Beachhead Effect." *The American Economic Review*, vol. 78, September, pp. 773–85.

Blanchard, Olivier, and Lawrence Summers. 1986. "Hysteresis and the European Unemployment Problem." In Stanley Fischer, ed., *NBER Macroeconomics Annual* 1986, pp. 15–78. Cambridge, MA: MIT Press.

Dixit, Avinash. 1989. "Hysteresis, Import Penetration, and the Exchange Rate Pass-Through." *Quarterly Journal of Economics*, vol. CIV, May, pp. 205–28.

Feenstra, Robert C., Joseph E. Gagnon, and Michael M. Knetter. 1996. "Market Share and Exchange Rate Pass-Through in World Automobile Trade." *Journal of International Economics*, vol. 40, pp. 187–207.

Froot, Kenneth A., and Paul D. Klemperer. 1989. "Exchange Rate Pass-Through When Market Share Matters." *The American Economic Review*, vol. 79, September, pp. 637–54.

Fuhrer, Jeffrey. 1995. "The Phillips Curve Is Alive and Well." Federal Reserve Bank of Boston, *New England Economic Review*, March/April, pp. 41–56.

Gagnon, Joseph E., and Michael M. Knetter. 1995. "Markup Adjustment and Exchange Rate Fluctuations: Evidence From Panel Data on Automobile Exports." *Journal of International Money and Finance*, vol. 14, no. 2, pp. 289–310.

Gordon, Robert J. 1977. "Can the Inflation of the 1970s Be Explained?" *Brookings Papers on Economic Activity*, vol. 8, pp. 253–77.

———. 1982. "Inflation, Flexible Exchange Rates, and the Natural Rate of Unemployment." In Neil Baily, ed., *Workers, Jobs, and Inflation*, pp. 89–158. Washington, DC: The Brookings Institution.

———. 1997. "The Time-Varying NAIRU and Its Implications for Economic Policy." *Journal of Economic Perspectives*, vol. 11, no. 1, pp. 11–32.

Motley, Brian. 1990. "Has There Been a Change in the Natural Rate of Unemployment?" Federal Reserve Bank of San Francisco, *Economic Review*, Winter (no. 1), pp. 3–16.

Sachs, Jeffrey D. 1986. "High Unemployment in Europe: Diagnosis and Policy Implications." NBER Working Paper no. 1830, February.

Staiger, Douglas, James H. Stock, and Mark W. Watson. 1997. "The NAIRU, Unemployment, and Monetary Policy." *Journal of Economic Perspectives*, vol. 11, no. 1, pp. 33–50.

Tootell, Geoffrey M.B. 1994. "Restructuring, the NAIRU, and the Phillips Curve." Federal Reserve Bank of Boston, *New England Economic Review*, September/October, pp. 31–44.

Weiner, Stuart E. 1993. "New Estimates of the Natural Rate of Unemployment." Federal Reserve Bank of Kansas City, *Economic Review*, vol. 78, no. 4, pp. 53–69.

Economic Profit and Performance Measurement in Banking

Ralph C. Kimball

Economist, Federal Reserve Bank of Boston, and Associate Professor of Finance, Olin Graduate School of Business, Babson College. *The author wishes to thank Richard Kopcke, John Jordan, and Eric Rosengren for helpful comments. The opinions expressed are those of the author and not necessarily those of the Federal Reserve Bank of Boston or the Federal Reserve System.*

Successful bank operation requires managers to weigh complex trade-offs between growth, return, and risk. In recent years banks increasingly have adopted innovative performance metrics such as risk-adjusted return on capital (RAROC) and economic value added (EVA^{SM})[1], which assist managers in making such difficult and complex decisions. These innovative measures all share as a basis the concept of economic profit, rather than accounting earnings. By forcing line managers to include the opportunity cost of equity when making investment and operating decisions, banks expect to elicit better decision-making by managers. By implementing performance measurement and incentive systems driven by economic profit and allocated equity capital, senior managers also hope to align managerial behavior more closely with the interests of shareholders.

This article analyzes the use of economic profit for measuring the performance of banks. In particular, since economic profit cannot be calculated without some imputation of equity, the article focuses on the allocation of equity capital to products, customers, and businesses. The first section of the article describes the use of economic profit to evaluate performance, to price transactions, and to reward managers. The second section describes in detail one performance measurement and incentive system, known as EVA^{SM}, which has been adopted by a considerable number of both banks and other companies. The third and fourth sections discuss the shortcomings of performance metrics founded on economic profit, which may distort banks' investment and operating decision-making. These metrics assume that it is possible to allocate earnings and equity capital to lines of business, products, and customers in a way that isolates the economic revenues and costs of each activity. However, if lines of business are related, either in the production of output or in their use of capital, then this isolation may not be possible, and these methods of measuring performance may mislead managers. The conclusion argues that banks need to recognize the ambiguities inherent in the calculation of economic profit and be prepared to create and apply multiple specialized performance metrics.

I. Economic Profit and Performance Measurement in Banks

Economists and accountants differ on the proper definition of profit. To the accountant, profit is the excess of revenues over expenses and taxes and is best measured by earnings. To the economist, earnings fails to include an important expense item, the opportunity cost of the equity capital contributed by the shareholders of the firm. A firm earns economic profits only to the extent that its earnings exceed the returns it might earn on other investments. Thus, earnings will always exceed economic profits, and a firm can be profitable in an accounting sense yet unprofitable in an economic sense.[1]

This conceptual difference has important practical implications. If managers attempt to maximize earnings (or growth of earnings) rather than economic profit, they will invest additional units of equity capital so long as the marginal contribution to earnings is positive. But if they do so, the marginal contribution of the last unit of equity capital will be zero and less than its opportunity cost, and the average return to equity capital may be greater or less than its opportunity cost depending upon how much equity is used. In contrast, a manager who maximizes economic profits will add units of equity capital only until the marginal contribution of capital is equal to its opportunity cost, and the average return to equity capital will equal or exceed its opportunity cost.

As a result, firms that make business decisions without explicitly incorporating the opportunity cost of equity will be inefficient users of equity capital, engaging in investment projects that generate low returns to shareholders.[2] In 1995, a year of robust earnings, one study estimated that fewer than half of the 1,000 largest industrial and nonfinancial firms earned sufficient returns to cover their opportunity cost of capital (see Ross 1997).

Banks and other companies have begun to address this issue by incorporating an explicit opportunity cost of equity into their decision processes. In particular, a number of banks have incorporated a measure of economic profit in three key areas: strategic decision-making, product pricing, and performance evaluation and incentive compensation.

[1] EVA[SM] is a registered servicemark of Stern Stewart & Co.

[2] While conventional capital budgeting models such as net present value or internal rate of return explicitly include a cost of equity capital, many decisions taken outside the capital budgeting process, such as product pricing or entry or exit from a particular line of business, may not.

Strategic Decision-Making

Businesses with different risk characteristics require different proportions of equity to achieve the same risk exposure. Evaluating businesses only on the level and rate of growth of their reported earnings fails to take into account differences in their use of equity, and the fact that shareholders may have different required rates of return reflecting the risk of the equity invested. Thus, when allocating scarce resources or when deciding to enter or exit a new line of business, managers compare a return on equity (ROE) for the business unit relative to an appropriate hurdle cost of equity. Business units earning an ROE in excess of a risk-adjusted opportunity cost of that equity are candidates to receive additional resources, while those earning less than this opportunity cost of equity are candidates for corrective action. In recent years, such calculations have been extended from lines of business to products, distribution channels, and even customers.

Pricing

As noted above, different products, customers, or transactions will absorb different amounts of equity capital, with larger and more risky transactions requiring more equity than smaller, less risky ones. To ensure that a transaction is profitable, managers must assign the appropriate amount of capital and a required contribution to equity must be calculated and incorporated in the price applied to the transaction. This use of allocated capital to ensure adequate pricing was first implemented by Banker's Trust in its RAROC system, which subsequently has been adopted by many other commercial banks.

In the RAROC system, the required rate on a loan comprises a cost of funds, a charge for non-interest expenses, a premium for credit risk, and a capital charge. The great contribution of the RAROC system was to include explicit charges for both the credit risk premium and the use of capital. By doing so, it ensures that banks price individual loans to cover credit risks and generate an adequate return for shareholders. An example of the use of the RAROC system to price loans is shown in Table 1. The capital charge is determined as the product of the proportion of equity capital assigned to support the loan and the required pre-tax hurdle rate on equity. As shown in Table 1, a loan rate of 11.25 percent will permit the bank to earn a 15 percent return on the equity required to back the loan. If the bank can obtain a rate greater than 11.25 percent, then it will earn an economic profit, while a

Table 1
The RAROC Method of Pricing Loans

Component	Example	Source
Funds Transfer Cost of Funds	5.45%	Funds Transfer Pricing System
Required Loan Loss Provision	1.25	Credit Risk Model
Direct Expense	.70	Customer/Product Cost Accounting System
Indirect Expense	.45	
Overhead	.40	
Total Charges before Capital Charge	8.25%	
Capital Charge	3.00	Allocated equity/loan = 12%
Total Required Loan Rate	11.25%	Opportunity cost of equity = 15% After tax capital charge = .12 x .15 = 1.80% Tax rate = .4 Pre-tax capital charge = 1.80%/.6 = 3.0%

loan rate between 8.25 and 11.25 percent generates positive earnings but an ROE of less than 15 percent.

Incentives

A relatively new but increasingly important area of study, the economics of organizations, analyzes the relationship between organizational structure and performance. One key concept in this new theory is that different agents (managers) within the organization have varied amounts of specific information concerning their businesses, products, and customers. An organization becomes more efficient by allowing investment and operational decisions to be made by those managers or groups of managers with the most specific knowledge concerning a particular decision. Thus, efficient use of specific information argues for a decentralization or devolution of decision-making to those line managers with the most information. Management innovations such as total quality management, quality circles, empowerment, and self-directed teams are all examples of the delegation of decision rights to line managers and employees to make more effective use of specific knowledge (See Wruck and Jensen (1997); Brickley, Smith, and Zimmerman (1997); and Bacidore, Boquist, Milbourn, and Thakor (1997)).

An obstacle that must be overcome, however, in such a decentralized system is the existence of agency costs. Agency costs occur when the interests of the managers are not necessarily the same as those of the shareholders of the firm, so that decentralization of decision-making to line managers can result in decisions that maximize the welfare of the managers rather than that of the shareholders. Such agency costs often appear in the form of higher costs, overinvestment, and suboptimal levels of risk incurred.

A performance measurement and incentives system that aligns more closely the interests of shareholders and managers can resolve this apparent conflict between the delegation of decision-making and agency costs. Most of these systems use some variation of economic profit to measure and reward the performance of managers. They take into account the opportunity cost of the equity capital that must be allocated out to the operating units. For example, in the EVA^{SM} system, discussed below in detail, managers' compensation depends upon either the level of or the increment to the economic value added of their particular unit. Because EVA^{SM} approximates economic profit, it reduces agency costs and permits greater decentralization in decision-making.

II. The EVA^{SM} Performance Measurement System

The EVA^{SM} performance measurement and incentive system was developed by the consulting firm Stern Stewart & Co. and it is the best known of a number of similar systems.[3] The EVA^{SM} system is built on the concept of economic value added, defined as the excess of adjusted earnings over the opportunity cost of the capital involved:

$$EVA = \text{Adjusted earnings} - c^*K$$

where earnings as defined by generally accepted accounting principles (GAAP) are adjusted to better represent economic earnings, c is the opportunity cost of equity, and K is the amount of equity used by the

[3] Other consulting firms' performance metrics based on the concept of economic profit include Holt Value Associates' cash flow return on investment (CFROI), Boston Consulting Group's total business return (TBR), and LEK/Alear Consulting Group's shareholder value added (SVA). See Myers (1996).

unit being measured.[4] EVA^{SM} can be calculated for the firm as a whole, but when used as a basis for an incentive system or to measure the performance of business units or individual managers, the earnings of and amount of equity capital used by these business units must be identified, so that their EVA^{SM} can be calculated.

Managers can improve the EVA^{SM} of their units in three ways: by increasing adjusted earnings, either through improved margins or additional sales; by reducing the equity capital used by the unit; or by reducing the cost of equity. However, measures taken to affect any variable in the EVA^{SM} equation will most likely affect the others, so that managers must take into account and manage trade-offs among the key variables. For example, earnings can be increased through expansion, but such expansion requires an increase in investment. EVA^{SM} will increase only to the extent the additional earnings generated by the expansion exceed the marginal cost of the additional equity capital involved. Similarly, a firm might increase EVA^{SM} by increasing its use of debt and decreasing the amount of equity (K) used. But as K decreases, the riskiness of the equity investment increases and c, the cost of equity, increases, so that EVA^{SM} will increase only if the percentage decline in K is greater than the percentage increase in c.

It is precisely because it requires them to manage these trade-offs at the margin that many managers believe that EVA^{SM} is superior to more conventional GAAP-based performance measures such as earnings or return on equity (ROE). As discussed above, a manager focused on maximizing earnings or the growth rate of earnings, without taking into account the opportunity cost of equity capital, will invest in new projects until the marginal contribution of the last project to earnings is zero. But if the marginal contribution of the last project is zero, then it is substantially less than the opportunity cost of capital at the margin, and the firm will be investing equity capital that the shareholders could better employ elsewhere. Such firms will grow and have positive GAAP earnings, but they will be inefficient users of equity and will fail to generate rewards for the shareholders as high as might be obtained in other uses.

Similarly, if managers focus on maximizing ROE, or the difference between ROE and some hurdle rate, then another problem appears. Logically, maximization of ROE requires that all projects except the one with the highest expected ROE be abandoned. For example, suppose a firm has three potential projects, with expected ROEs of 30, 25, and 20 percent, respectively. The opportunity cost of equity capital is 15 percent. A manager maximizing ROE or the difference between ROE and the opportunity cost of equity capital will pick only the first project, despite the fact that the other two would also generate economic profits for the firm. Thus, a firm that uses a performance metric based on ROE will tend to underinvest and grow more slowly than it should.

A firm using EVA^{SM} would avoid either of these outcomes because managers would be forced to internalize the trade-off between growth and the return to additional equity. A manager maximizing EVA^{SM} would invest until the last project generated an ROE just equal to the opportunity cost of the equity capital employed. Growth would be pursued but only so long as additional projects enhance economic profit. Proponents argue that in addition to causing managers to economize on their use of equity capital, the most expensive part of the firm's balance sheet, an EVA^{SM}-based system makes explicit each manager's contribution to economic profit, and by doing so results in increased focus and commitment.[5]

One problem with any incentive compensation system (and an illustration of agency costs) is that it can be manipulated by managers to maximize their compensation without necessarily increasing the profits of the firm. For example, if the incentive compensation system considers only the manager's performance this period, then it is often possible for a manager to take actions that raise reported perfor-

[4] EVA^{SM} can be defined with respect to either total assets or equity. In a total assets formulation, adjusted earnings represents net operating earnings after tax but before interest expense, while c represents the weighted average cost of capital, and K the total assets of the firm. In the equity formulation, used in the rest of this article, adjusted earnings represents net income after interest and taxes, c is the cost of equity, and K is equity capital.

Stern Stewart has identified over 160 potential adjustments to the GAAP definition of net income that it believes result in a better reflection of economic earnings. For banks, there are four major adjustments: using actual net charge-offs rather than the loan provision, using cash taxes rather than the tax provision, excluding securities gains and losses, and considering nonrecurring events as an adjustment either to earnings or capital, on a case-by-case basis. See Uyemura, Kantor, and Pettit (1996).

[5] Proponents of EVA^{SM} also argue that it is more closely correlated with return to shareholders than are traditional GAAP accounting measures such as earnings, return on assets, or ROE. Such assertions have been criticized. For the proponents' arguments see Stewart (1991); Stern, Stewart and Chew (1995); Uyemura, Kantor, and Pettit (1996); and O'Byrne (1997). For the opposite view see Kramer and Pushner (1997) and Lehn and Makhija (1996).

mance this period but depress it in succeeding periods.[6] For example, in banking, current-period operating earnings can be enhanced by cutting service levels and relaxing credit standards, but such actions will depress future operating earnings as disgruntled customers switch to competitors and credit losses increase. Indeed, incentive systems resemble tax systems in that one of the biggest challenges in designing such systems is to identify and close loopholes that facilitate gaming.

As with any other incentive system that focuses solely on current-period performance, an EVA^{SM}-based system can be manipulated to maximize current incentive compensation at the cost of future reported performance. In many firms, this time horizon problem is addressed by creating an incentive compensation account for each manager into which both positive and negative annual payments are made. Managers are permitted to withdraw only a maximum percentage of the balance in the incentive account in any one year.[7] By creating a rolling five-year time horizon for the effective vesting of the incentive compensation, the manager's incentives to manipulate short-term performance at the cost of long-term performance is limited, since any increase in this period's incentive compensation would be offset by negative incentive compensation in the succeeding periods.[8]

III. Related Operations and EVA^{SM}

To be effective in reducing agency costs and facilitating the devolution of decision-making, any performance measurement and incentive system must apply not just to senior management, but also at the divisional, product, and customer levels. Only by application at the business unit level can a performance measurement system be expected to affect the behavior of managers at these levels. However, application of any measurement and incentive system based on economic profit, whether EVA^{SM} or another, to subunits of a bank is based on a key assumption: that it is possible to isolate the earnings contribution of each business unit of the bank and the proportion of the bank's equity capital it uses. In effect, calculation of economic profit at the business unit level views the firm as being the aggregation of individual units, and

Lines of business, divisions, products, or other subunits are related operationally when the level of activity in one unit affects the earnings of another. Relatedness can affect revenues as well as expenses.

the earnings and equity capital of the firm as being the sum of the individual earnings and equity capital used by the subunits. But this "the whole is the sum of the parts" assumption may not be valid if either the earnings of one unit are affected by the actions of another or the economic risks faced by different units are imperfectly correlated. This section discusses the effects of related operations upon the calculation of economic profit, while Section IV discusses issues associated with the allocation of equity capital to business units.

Lines of business, divisions, products, or other subunits are related operationally when the level of activity in one unit affects the earnings of another. An extreme example of related operations is the production of joint products, where a process results in the production of two separate products in fixed relative proportions. A classic example of joint production used in many textbooks is the slaughter of a steer, resulting in both beef and leather. Neither product can be produced without the other, and the volume of each is more or less fixed with respect to the volume of the other. But operations can be related in many circumstances other than strict joint production. In many situations business units share common expense bases, products, distribution channels, or customers.

[6] When, presumably, the manager would no longer be with the firm.

[7] For example, suppose a manager earns an incentive compensation bonus in 1997 of $10,000. The $10,000 is deposited as deferred compensation into an incentive compensation account. The manager is allowed to withdraw only 20 percent of the account in any one year, so that a maximum of $2,000 can be withdrawn in 1997 with a balance of $8,000 carried over to the next year. If in 1998 the manager earns another $5,000 in incentive compensation, the balance in the account will be $13,000, and the manager may withdraw a maximum of 20 percent, or $2,600. On the other hand, if the manager's unit does poorly and the manager earns an incentive compensation payment of a negative $5,000, then the balance in the account declines to $3,000 and the manager may withdraw only $600.

[8] Such systems work only if the manager is willing to accept negative incentive compensation in poor years, and if any remaining balance in the incentive compensation account is forfeited if the employee leaves the firm. The latter condition also acts as "golden handcuffs" to reduce turnover of key managers.

For example, bank products such as credit cards and home equity loans may share the same revolving loan system used to process account payments and statements. Similarly, advertising that stresses a bank's willingness to lend may affect more than one loan product.

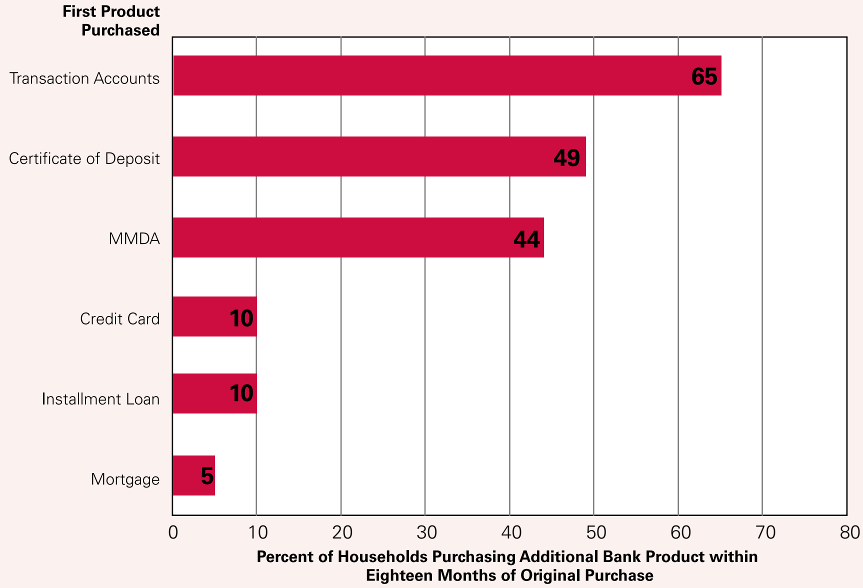
The existence of shared expense bases means that these costs must be divided among the subunits that share them, if the earnings of each are to be calculated. If the expenses of the shared cost center vary directly with volumes, they can be allocated to the subunits in proportion to their usage. But in most cases, the expenses of the cost center are relatively fixed or vary less than proportionately with volumes. Cost allocations then become arbitrary and can lose their economic usefulness. Consider the example of a bank operating a revolving loan system used by two loan products: credit cards and home equity loans. An economist would argue that each product or transaction should be charged its marginal cost. But in the case of the revolving loan system, most of the costs are fixed in the form of system development and maintenance, and the cost of executing an additional transaction or adding a product is almost zero. Thus, the manager for either product could argue that it should not be allocated any of the costs of the system since the marginal cost of adding the product or transaction to the system, once it exists, is zero. In reality, the costs of such systems are usually allocated on the basis of usage, so that each product is effectively charged an average cost per transaction times the number of transactions executed.

Even if one is willing to overlook the distortions introduced to decision-making by the use of average costs rather than marginal costs, one is still left with the result that changes in the volume of activity of one product will affect the costs of the other. For example, in the case of the revolving loan system, should the credit card product increase its volume while the home equity product did not, then the fixed costs of the revolving loan system would be spread out over a larger number of transactions, and the average cost per transaction would fall. If both products are charged the equivalent of average transaction cost times the number of transactions, then the allocated expenses of the home equity product will fall, solely because it shares an information system that has economies of scale.[9]

Relatedness can affect revenues as well as expenses. For example, to the extent that advertising is positively correlated with sales, all the business units

Figure 1

Propensity to Purchase Additional Products Based on First Product Purchased



Source: Gemini Consulting.

[9] Management accountants try to mitigate this relationship between volume variances and allocated expenses by allocating expenses on the basis of budgeted volumes rather than actual ones. That is, the expenses of the revolving loan system would be allocated on the volumes the credit card and home equity products expect to occur rather than actual ones. Thus, unexpected increases or decreases in volumes do not affect the amount allocated. However, this approach is a short-term remedy at best. While an unexpected increase in the volumes of the credit card business will not affect the costs allocated to the home equity product this period, next period the credit card business will revise its planned volumes upward to match the actuals, and at that point the proportion of systems costs allocated to the home equity product will decline.

may benefit from a corporate advertising campaign. Similarly, to the extent that customers prefer to purchase on a relationship basis and cluster their product purchases with one supplier, the acquisition of a new customer by one unit may enhance the revenues of other units. This is illustrated in Figure 1, which compares the propensity of the retail customers of a large money center bank to purchase a second product within 18 months of purchasing different initial products. As the figure shows, 65 percent of customers who opened a transactions account as their first product purchased a second product within 18 months, compared to only 5 percent of customers who took out a mortgage as their first product. Thus, it would appear that the initial sale of a transactions account has positive externalities for other retail products and that this positive impact is much larger than those connected with the initial sale of other products.

In such cases the existence of shared expenses or revenues can make it impossible to isolate costs or revenues in an economically meaningful way. In the extreme case of joint production, the sale of the two products generates two streams of marginal revenue, but there is only a single shared marginal cost. As a result, economists have long recognized that it is impossible to meaningfully calculate the profitability of either of the joint products, and that the condition for profitability maximization is the production of the joint products until the sum of their marginal revenues just equals the joint marginal cost. In the case of relatedness arising through sharing, something similar occurs. Ideally, the revenues and expenses allocated to a business unit would consist not only of those that can be directly traced to a change in the volumes of that unit, but also of the incremental revenues and expenses of other subunits that result from the change in volumes of the first unit. Banks often attempt to accomplish this by implementing transfer pricing systems. Thus, a retail branch that serves corporate customers of the middle market group may receive an internal credit to cover the associated incremental expenses. But relatedness often appears in subtle and intangible ways, so that it is unlikely that transfer pricing can capture all of the effects.

Relatedness would not be an issue in performance measurement if the degree of relatedness was small or if positive and negative effects for each unit canceled each other out. While no empirical data exist that would permit us to measure the effects of related operations on reported revenues or expenses, intuitively one can expect relatedness to exist and to increase in importance, the smaller the subunit being considered. Moreover, there is reason to believe that the effects of relatedness are not unbiased, but instead act to cause some subunits to systematically underestimate their contribution to earnings and others to overestimate it.

Indeed, some form of related operations is a necessary condition for different lines of business to exist or different products to be produced in the same firm. If no relatedness is present between lines of business or products, then each business or product could operate independently with no loss in value, and there is no economic rationale for joining them in the same firm. Increasingly, this argument is being accepted by managers, as demonstrated by the increased number of spin-offs and sales of "nonstrategic" businesses or products. It is only when benefits exist from joint operation that lines of business or products should be combined in one firm, so that these benefits can be captured.[10] In effect, multidivisional or multi-product firms exist because relatedness causes the value of the whole to be more than the sum of the value of the parts (see Zimmerman 1997).

Where relatedness exists, any performance metric that is calculated only on the allocated revenues and expenses of a single business unit, such as EVA^{SM}, will be an inaccurate measure of that unit's contribution. The contributions of business units that generate negative expense or positive revenue effects for other units will be underreported, while the contributions of subunits that enjoy either lower expenses or higher revenues as a result of the activities of others will be exaggerated. Managers attempting to maximize unit EVA^{SM} will underinvest in units that generate positive externalities and overinvest in units that receive them. The failure to incorporate relatedness into the calculation of EVA^{SM} leads to a "management myopia" where each manager is trying to maximize business unit EVA^{SM} but not bankwide EVA^{SM}. As discussed in the box, "Relatedness and Incentive Systems," incentive systems can be constructed to encourage managers to take into account the effects their decisions will have on other business units, but such incentive systems are complex and usually are only partially effective.

[10] In the strategic planning literature the effect of relatedness is captured in the concepts of core competencies and horizontal strategies. A core competency is a skill or activity that cuts across lines of businesses or products and is the basis for the competitive advantage of the firm. A horizontal strategy is one that is built around a core competency. See Prahalad and Hamel (1990) and Porter (1985), Chapters 9–11.

Relatedness and Incentive Systems

In cases where subunits generate substantial externalities, implementation of an incentive compensation system based on subunit profitability can lead to perverse results by encouraging managers to ignore the effects of their actions on other subunits. This adverse effect has long been recognized, and a variety of approaches have been developed to address it. They can be summarized as linked incentives, hierarchical grouping, and hybrid systems.

Linked Incentives

In a linked incentives approach, the incentive compensation of a manager is determined by two or more components: the performance of the manager's own unit, and the performance of either the firm as a whole or some other subunit. For example, 70 percent of a manager's incentive compensation might be determined by the EVA^{SM} of her own unit, and 30 percent by the EVA^{SM} of the bank as a whole or by the EVA^{SM} of some other related unit. The intent is to cause the manager to optimize the trade-offs between the EVA^{SM} of her own subunit and that of other parts of the bank.

Such linked incentive plans obviously expand the manager's horizon but they have two potential drawbacks. First, their efficacy in optimizing decision-making requires that the effects of relatedness be accurately identified and quantified in the incentive scheme. If they are not, the incentive scheme will not elicit optimal decision-making.[a] A second drawback of cross-linked schemes is that the more effective they are in identifying and quantifying externalities, the more complex they become. If the level of activity in one business unit affects multiple other units, any incentive scheme that accurately reflects this becomes immensely complex and unwieldy. Even if the average impact of one unit's activities on the rest can be determined, the marginal impact may vary from transaction to transaction, so that any simple cross-linkage scheme will fail to elicit efficient decision-making.

Hierarchical Grouping

A second approach to internalizing relatedness is through hierarchical grouping. In a grouping approach, the manager of each business unit is responsible for and compensated only on the EVA^{SM} of that unit. Related business units are then combined in groups, with the manager of the group evaluated and compensated on the EVA^{SM} of the group. The manager of the group thus has an incentive to maximize the synergies that might exist between the subunits. This approach avoids the rigidity of the cross-linkage approach, since the group manager's decisions are made on a case-by-case basis rather than with respect to an oversimplified and probably inaccurate formula. However, while the manager of the group has an incentive to maximize the cross-unit synergies, doing so could place him in conflict with the subunit managers, especially if the actions required to maximize group EVA^{SM} do not maximize business unit EVA^{SM}. To avoid this divergence in interests between unit manager and the group manager, business unit managers often are penalized if the group does not meet its EVA^{SM} goals.[b]

[a]For example, suppose a manager is operating under an incentive scheme where she will receive seven-tenths of 1 percent of the incremental EVA^{SM} of her own unit and three-tenths of 1 percent of the incremental EVA of the rest of the firm. The manager is evaluating an investment that will decrease her unit's EVA by $100,000 but will increase the EVA^{SM}of other units by $200,000. Although acceptance of the project would increase the firm's EVA^{SM} by $100,000, the manager has an incentive to reject it. If she accepts the project, her incentive compensation will actually decline by $100.

A drawback to grouping is that while it addresses synergies that occur among the subunits of the group, it does not address those that occur among different groups. To address these higher-level synergies, divisions consisting of related groups must be created, with division managers evaluated and compensated on the EVA^{SM} of the division. But this solution, in turn, does not address the issue of synergies across divisions. In effect, grouping is probably a more flexible way to address the issues of related operations, but it requires the creation of a hierarchical management structure. As more layers of management are added, the hierarchical structure created has its own drawbacks in the form of higher costs, slower decision-making, and a reversal of the decentralized decision-making that was the original objective of implementing an EVA^{SM}-based system.

Hybrid Systems

Yet another approach is to expand the performance measurement system to include nonfinancial variables. For example the "balanced scorecard" approach measures managers in areas such as "leadership," "customers," and "people," as well as the more traditional financial goals. Such a scheme can address the issue of relatedness by including key operating measures that affect other business units. These key operating measures might be the number of new customers added or the number of leads generated for other areas of the bank. For example, if, as shown earlier in Figure 1, the origination of new transactions accounts has positive externalities for other areas of the bank, then a manager in the retail banking business might be evaluated both on the EVA^{SM} of his unit and also on the number of new transactions accounts originated.

Of course, such hybrid systems of financial and nonfinancial variables come with their own set of drawbacks. In particular, by evaluating a manager's performance on an "apples and oranges" basis, the system loses both its objectivity and its capacity for internalizing trade-offs. For example, suppose a manager has the dual objectives of generating 1,000 new transactions accounts and an EVA^{SM} of $100,000. Unfortunately, such a system gives the manager little guidance about how to make decisions at the margin. For example, is it better to generate only 900 transactions accounts but an EVA^{SM} of $110,000, or 1,100 transactions accounts and an EVA^{SM} of only $90,000?

No incentive compensation system is perfect, and many firms and banks end up using a combination of systems.

While the efficacy of incentive compensation systems in encouraging managers to capture the effect of cross-unit synergies can be increased in a number of ways, none is perfect and most involve costs of their own. Many firms and banks end up using a combination of hierarchical groupings, hybrid performance measurement systems, and linked incentives to address this issue. Such integrated systems must be carefully constructed and monitored to ensure that they have a positive effect on the overall performance of the bank.

[b]This approach does not completely avoid conflict. Subunit managers have an incentive to cooperate with the group manager until the group goal is achieved. At that point the conflict between the subunit manager's objective to maximize subunit EVA^{SM} and the group manager's objective to maximize group EVA^{SM} reemerges.

Table 2
Peer Group Approach to Allocating Equity Capital for Consolidated Amalgamated Bank

Line of Business	(1) Assets ($millions)	(2) Equity ($millions)	(3) Equity/Assets (Percent)	(4) Return on Assets (Percent)	(5) σ_{ROA}	(6) Z-Ratio
Credit Cards	20,261	2,018	9.96	4.94	1.08	13.80
Mortgage Banking	11,314	1,949	17.23	4.96	2.78	7.98
Subprime Lending	5,072	1,666	32.77	14.67	7.96	5.96
Total	36,647	5,633	15.37	5.99	1.29	16.56

Source: Compustat and author's calculations.

IV. Allocating Equity Capital

If some variation of economic profit is to be calculated at the business unit level, then the bank's equity capital, as well as its earnings, must be disaggregated and divided among the business units. This allocation of equity capital is critical, since without it the opportunity cost of equity cannot be calculated. In any firm, equity has two different functions: as a source of funding to purchase equipment, premises, and inventory, and as a cushion to protect debt holders against loss in the event of operating losses. Because banks hold relatively few of their assets in the form of real assets, equity's function as a cushion for economic risk is especially important in banking. The proportion of equity needed to support a line of business, product, or customer within the bank will depend upon the riskiness of the activity, with riskier activities requiring additional capital.[11] Thus, the amount of equity capital allocated to a particular business unit will depend both on the scale of operations (for example, the amount of assets held) and the riskiness, so that a small but risky subunit could require as much equity as a large but low-risk one.

Stand-Alone Allocation Methods

One approach to allocating equity bases the allocations on the capital structure of independent "pure play" peers. To do so, a bank would construct for each line of business a group of publicly traded peers and allocate capital according to the average capital ratio of the peer group.[12] For example, the mortgage banking business would be assigned equity as though it were an average, independent, publicly traded mortgage banker. While this approach has the advantage of being based on objective market data, actual implementation quickly reveals several drawbacks. The number of independent publicly traded peers may be small or in some cases nonexistent, and these peers may differ in important respects from the business being analyzed. And even if a sufficient number of publicly traded peers exist, their capital ratios may vary significantly, so that management must choose among a possible range of capital allocations rather than a closely clustered point estimate.

This approach is illustrated for a fictional Consolidated Amalgamated Bank in Table 2. The Consolidated Amalgamated Bank is constructed from data for three separate publicly traded monoline lenders: a mortgage banker, a credit card bank, and a subprime consumer lender.[13] In Table 2, the capital allocated under the peer group approach is assumed to be the same as the units' actual equity capital in their true identity as publicly traded independent firms. As shown there, while the bank as a whole has an equity-to-asset ratio of about 15 percent, the equity-to-asset ratios of the individual businesses vary from about 10 percent for the credit card business to 33 percent for the subprime lending business.

While the peer group method of allocation clearly differentiates among the lines of business in terms of the amount of capital allocated, it does not necessarily result in equal probabilities of insolvency across different lines of business.[14] For example, if consumer finance companies have on average a higher probability of insolvency than do mortgage banks, then allocation of equity capital based on the average of their respective capital structures will result in a higher

[11] Riskiness is usually measured as the volatility of returns, for example, the standard deviation of the return on assets.

[12] Allocations to products and customers would usually reflect the line of business to which they belong.

[13] This approach was necessary because no bank publishes line-of-business results on a quarterly basis over a sufficient time period to permit calculation of expected returns and their covariance.

[14] Insolvency for a line of business should be interpreted as the probability that the losses of the line of business will exceed the equity capital allocated to it.

Table 3
Capital Allocations for Consolidated Amalgamated Bank with Equal Probability of Insolvency

Line of Business	(1) ROA (Percent)	(2) σ_{ROA}	(3) Z*-Ratio	(4) Equity/Assets (Percent)	(5) Equity ($millions)
Credit Cards	4.94	1.08	13.80	9.96	2,018
Mortgage Banking	4.96	2.78	13.80	33.40	3,779
Subprime Lending	14.67	7.96	13.80	95.18	4,827
Total Bank	5.99	1.29	27.12	28.99	10,624

Required equity capital for bank to achieve Z* = 13.80:
K = (13.80)(1.29) − 5.99 = 11.81%
Equity capital = (11.81%)(36,647) = $4,329 million.

Source: Columns 1, 2, and 3: Table 2 and author's calculations. Column 4: (Column 3 × Column 2) − Column 1.

probability of insolvency for the bank's consumer lending business than for its mortgage origination business. One index of the probability of insolvency is the Z-ratio,[15] defined as:

$$Z = (ROA^* + K)/\sigma_{ROA} \qquad (1)$$

where

ROA* = the pretax expected return on assets, usually defined as the historical mean ROA,

K = the ratio of equity capital to assets, and

σ_{ROA} = the standard deviation of ROA.

Thus, the Z-ratio is a function of the normal profit margin of the bank, the variation in that profit margin, and the equity capital available to absorb that variation. In effect, the Z-ratio measures the number of standard deviations by which ROA would have to decline before the book equity capital of the bank would be exhausted. The relationship between the Z-ratio and the probability of insolvency is an inverse one, with higher Z-ratios indicating a lower probability of insolvency.[16] The last four columns of Table 2 calculate the Z-ratio for each line of business and for the bank as a whole. As shown there, the Z-ratios differ significantly across the lines of business, with the credit card business having a substantially lower probability of exhausting its assigned equity than do the mortgage banking and subprime lending businesses.

An alternative approach allocates equity capital based on each business's cash flow so as to create an equal probability of insolvency. Equation (1) above can be rewritten to express the capital-to-asset ratio required to achieve a given target Z-ratio, as follows:

$$K^* = Z^*\sigma_{ROA} - ROA^* \qquad (2)$$

where K* is the required capital-to-asset ratio to achieve a target Z-ratio equal to Z*. In this approach each line of business will be allocated capital until its Z-ratio equals Z*. Application of this approach to Consolidated Amalgamated is illustrated in Table 3, which assumes that each line of business is allocated capital to achieve a Z-ratio of 13.8, the initial Z-ratio of the credit card business. This approach results in substantially higher equity-to-asset ratios for the mortgage banking and subprime lending businesses. Indeed, the equity capital-to-asset ratio of the subprime lending business increases from about 33 percent under the peer-group method to about 95 percent under the equal probability of insolvency approach. Similarly, if the required equity of the bank as a whole is the sum of the required equity for each of the lines of business, then the bank will require almost 89 percent more equity under the equal probability of insolvency approach than under the peer group approach.

[15] This measure was developed by Hannan and Hanweck (1988). Although Hannan and Hanweck called the risk index "g," in subsequent work it has generally been called "Z."

[16] If the assumption is made that the potential ROAs of the business are normally distributed, then the one-period probability of insolvency can be calculated as a function of the Z-ratio:

$$p = 1/[2Z^2]$$

However, empirical studies indicate that ROAs are not normally distributed, but instead are "fat-tailed," so that the actual probability of insolvency may be greater than that calculated using the assumption of normality. Moreover, this one-period probability may understate the true probability of insolvency because it measures the risk of a single-period loss being so large it wipes out equity. In reality, insolvency often occurs after a sequence of smaller losses occurring over several periods, indicating that serial correlation between negative shocks may exist.

Table 4
Capital Allocation for Consolidated Amalgamated Bank with Equal Probability of Insolvency and Diversification Effects

Line of Business	(1) Stand-Alone Equity ($millions)	(2) Diversification Effect	(3) Equity Allocation with Diversification Effect ($millions)
Credit Card	2,018	.4074	822
Mortgage Banking	3,779	.4074	1,540
Subprime Lending	4,827	.4074	1,967
Total Bank	10,624	.4074	4,329

Source: Column 1: Table 3. Column 2: $4,329 (from Table 3) ÷ $10,624 (from Table 3). Column 3: Column 1 × Column 2.

Allowing for Diversification

A comparison of the Z-ratios for the bank as a whole with the Z-ratios for the individual lines of business, as shown in Tables 2 or 3, reveals a drawback to both of these stand-alone methods of allocating capital. The Z-ratio for the bank as a whole is considerably greater than the Z-ratio for any of the three lines of business, indicating that the probability of insolvency for the bank is less than that of any of the lines of business. This occurs because the correlation in the ROAs of the individual businesses is less than perfect. To the extent such correlations are less than perfect, they will tend to dampen the fluctuations in returns for the bank as a whole, so that the risk of the bank will be less than the weighted sum of the risks of the individual businesses. In effect, the business units act as partial natural hedges for each other, reducing the need for equity capital. Thus, a bank with a diversified portfolio requires less equity capital to achieve any given probability of insolvency than do the business units on an aggregated stand-alone basis. This is shown at the bottom of Table 3, where the amount of equity capital needed for the bank as a whole to achieve a Z-ratio of 13.8 is calculated to be only $4.3 billion, less than half of the $10.6 billion calculated as the sum of the stand-alone allocations to the individual businesses.

Thus, in those situations where the ROAs of the individual businesses are imperfectly correlated, a discrepancy will result between the sum of the individual equity allocations to the different lines of business and the equity capital required when the effects of diversification are incorporated. This discrepancy creates obstacles to the evaluation of businesses and their managers. Ultimately, the larger the capital allocation, the more difficult it is for a line of business to earn an economic profit. If capital allocations to individual businesses exceed the actual capital of the bank, then managers may believe this "ghost capital" unfairly biases downward the reported return on equity of each business. The excess allocated capital can also create strategic issues, since the reported EVA^{SM}s of the business units will not sum to the EVA^{SM} of the bank. Theoretically it would be possible for each line of business to fail to earn its required opportunity cost of stand-alone equity, while the bank as a whole surpassed its required opportunity cost of equity based on actual equity capital, which includes the effects of diversification. In extreme cases, a bank might choose to exit a business based on an insufficient return to equity earned on allocated capital, when the return on equity on actual capital might be quite satisfactory.

Proportional Scaling

This problem can be addressed in two ways. The simplest is to scale back the allocations to the individual businesses so that the sum of the allocations equals the actual (diversified) capital of the bank. Thus, if the sum of the individual allocations is 200 percent of the actual capital of the bank, each allocation is reduced by one-half to make the sum of the individual allocations equal to actual capital. This approach is illustrated for Consolidated Amalgamated in Table 4, assuming that each line of business has the same probability of insolvency (from Table 3) and that the bank as a whole has a target Z-ratio of 13.8. In effect, this approach spreads the reduction in equity capital due to diversification across the lines of business in proportion to their initial stand-alone capital allocations.

Table 5

Effect of Scaled Reductions in Capital Allocations on Reported Economic Profit

Business Unit	Adjusted Earnings	Opportunity Cost of Allocated Capital before Diversification Effects	Reported Economic Profit before Diversification Effects	Opportunity Cost of Allocated Capital after 50% Diversification Effects	Reported Economic Profit after Diversification Effects	Incremental Economic Profit due to Scaled Reductions in Capital Allocations
A	100	100	0	50	50	50
B	100	70	30	35	65	35
C	100	50	50	25	75	25

Source: Author's calculations.

While simple to implement, this approach to incorporating the effects of diversification has serious conceptual drawbacks. By allocating the reductions in equity capital in proportion to the initial stand-alone capital allocations, inefficient users of capital receive a disproportionate increment to their economic profits. An example of this is shown in Table 5, which compares three lines of business before and after the scaled reductions in stand-alone allocations.[17] All three lines of business have the same adjusted earnings, but they differ in the amount of capital used and thus in their reported economic profits. If stand-alone capital allocations are scaled back by 50 percent to reflect the benefits of diversification, then the incremental effect on the reported economic profits of Business A, the most inefficient user of capital, will be double that of Business C, the most efficient user of equity capital. As a result, the simple scaling approach obscures the ability of senior management to distinguish among the business units in their efficiency in using equity capital.

Moreover, when the benefits from diversification are allocated in proportion to their initial stand-alone capital allocations, they are being allocated in proportion to the stand-alone total risk of each line of business, weighted by the dollar assets of each business. But the contribution of a particular line of business to the total risk of the bank will depend not only on the stand-alone risk of that line of business, but also on the correlations in returns among the different lines of business of the bank. A line of business with a low or negative correlation of returns with the other parts of the bank will diversify away more risk than will a line of business with a high positive correlation. A simple proportional reduction in stand-alone capital tends to over-allocate capital to lines of business units with low or negative correlations, and to under-allocate equity capital to business units with high positive correlations.

[17] The lines of business shown in Table 5 are fictional and are not those shown for Consolidated Amalgamated in Tables 2, 3, 4, 6 and 7.

Internal Betas

A second possible alternative to incorporating the effects of diversification in allocating capital is based upon the concept of "internal betas." In this approach, the relative risk contribution of each line of business is calculated as an internal beta, defined as the ratio of the covariance between the business unit's and bank's returns to the variance of the bank's returns:

$$\beta_{Bus} = \text{cov}(R_{bus,}\ R_{bank})/\sigma^2_{Bank} = (\sigma_{Bus}/\sigma_{Bank})\rho_{Bus,Bank}$$

where σ_{Bus} and σ_{Bank} are the standard deviations of the ROAs of the business unit and the bank as a whole, respectively, and $\rho_{Bus,Bank}$ is the coefficient of correlation of returns between the business and the bank. In this formulation the risk contribution of each business will depend on two factors, its stand-alone risk relative to the bank as a whole ($\sigma_{Bus}/\sigma_{Bank}$) and the degree of correlation between the returns of the business and the bank ($\rho_{Bus,Bank}$). The effect of the correlation in returns is unambiguous—the greater the correlation, the greater the risk contribution of the business—but the effect of the stand-alone risk of the business will depend on the sign of the correlation coefficient. If the correlation between the unit's and the bank's returns is positive, then the risk contribution of the business will increase in proportion to its stand-alone risk, but if the correlation in returns is negative, then the risk contribution of the business will decrease as the stand-alone risk of the business increases. Intuitively, if returns are negatively correlated, then variations in

Table 6
Allocation of Equity Capital for Consolidated Amalgamated Bank Using Internal Betas

Business Unit	(1) Standard Deviation of Returns (σ_{ROA})	(2) Correlation Coefficient ($\rho_{Bus,\ Bank}$)	(3) Internal Beta (β_{Bus})	(4) Equity Capital Ratio of Business (Percent)	(5) Allocated Equity Capital ($millions)
Credit Card	1.08	.762	.638	7.54	1,526
Mortgage Banking	2.78	.423	.911	10.77	1,217
Subprime Lending	7.96	.429	2.65	31.27	1,586
Bank Total	1.29				4,329

Source: Column 1: Table 2
Column 2: Compustat, author's calculations.
Column 3: (Column 1 ÷ 1.29) x Column 2.
Column 4: Column 3 x (4,329/36,647) (from Tables 4 and 2).
Column 5: Column 4 x Column 1, Table 2.

returns from the business tend to offset variations in returns on the bank as a whole, and the greater the variation in returns on the business (σ_{Bus}), the greater the reduction in the overall risk of the bank.

In the internal beta approach, the equity capital-to-asset ratio for each business unit is equal to the product of the unit's internal beta and the bank's overall equity capital ratio:

$$K_{Bus} = \beta_{Bus} K_{Bank}$$

where K_{Bus} is the capital-to-asset ratio of the business, β_{Bus} is the internal beta of the business, and K_{Bank} is the capital-to-asset ratio of the bank, including diversification effects. This approach is illustrated for the Consolidated Amalgamated Bank in Table 6. As can be seen there, the capital allocations under this approach differ substantially from the equal-scaling approach shown in Table 4. In particular, the business units with relatively low correlation in returns (mortgage banking and subprime lending) are allocated substantially less equity capital under this approach than the business unit (credit card) with a relatively high correlation in returns.

Marginal Capital

While the internal beta approach integrates both the stand-alone risk of the business and its interaction with the rest of the bank, its use to calculate the risk contribution of a business unit involves several restrictive assumptions. As discussed in the accompanying box, the internal beta approach measures the risk contribution of a business unit under the assumptions that the business already exists within the bank and that the relative size of the business (and of the other businesses in the bank) does not change. This means that the internal beta approach is most appropriate in a relatively static situation and results in biased allocations in more dynamic situations such as acquisitions or divestitures, or where business units are growing at different rates. Thus, in situations where the mix of businesses is changing, as a result of either strategic decisions or differential growth rates, capital should be allocated based on the business's marginal risk contribution.

Marginal capital can be defined as the incremental capital (for the bank as a whole) resulting from a change in the scale of operation of a business unit, assuming the probability of insolvency remains constant. For an acquisition or divestiture, marginal capital is measured as the difference between the required equity capital for the bank as a whole, including the business being bought or sold, and the required equity capital for the bank without the line of business. For an existing business that is expanding its scale of operations, it can be measured as the incremental capital for the bank as a whole associated with the incremental increase in volumes.

Marginal capital for each of the lines of business of Consolidated Amalgamated Bank is shown in Table 7 under the assumption that each line of business is being divested. That is, marginal capital is calculated as the difference in the bank's required capital, with and without the line of business in question. As can be seen in Table 7, marginal capital depends both on the extent of the correlation in returns between the business units in question and on the effect of the change on the diversification of the bank.

Adding a business that has a low positive correlation with existing businesses will require less incremental capital for the bank than will acquiring one with a high positive correlation, and acquiring a

Table 7

Calculation of Marginal Equity Capital for Consolidated Amalgamated Bank

Business Unit	(1) Required Equity Capital for Bank with All Three Business Units ($millions)	(2) Required Capital Ratio for Bank without Business Unit (Percent)	(3) Bank Assets without Business Unit ($millions)	(4) Required Equity Capital for Bank without Business Unit ($millions)	(5) Marginal Equity Capital ($millions)	(6) Marginal Capital Ratio (Percent)
Credit Card	4,329	21.74	16,386	3,562	767	3.78
Mortgage Banking	4,329	19.78	25,333	5,012	(683)	(6.04)
Subprime Lending	4,329	12.55	31,575	3,961	368	7.25
Total Allocated Capital					452	
Unallocated Capital					3,877	
Total Bank Capital					4,329	

Source: Column 1: Table 3.
Column 2: Author's calculations, using method from Table 3.
Column 3: Compustat.
Column 4: Column 2 × Column 3.
Column 5: Column 1—Column 4.
Column 6: Column 5 ÷ Column 1, Table 2.

business with a negative correlation with existing businesses can actually reduce the required capital, resulting in negative marginal capital. This is shown in Table 7 for the mortgage banking business. Because the correlation in returns between the mortgage banking business and the subprime lending business is negative (−0.53), adding the mortgage banking business to an existing combination of the credit card and subprime lending businesses actually dampens the variation in the aggregate and therefore reduces the required capital. Moreover, marginal capital is not constant but will vary as the size of the business in question varies relative to the size of the other businesses in the bank. As discussed in the box, "Internal Betas and Marginal Capital," the marginal capital associated with a given increment in the size of a business increases as the business unit becomes a larger proportion of the bank.

Capital Allocations and EVA^{SM}

Table 8 summarizes the results of Tables 2, 3, 4, 6, and 7 and shows the equity capital allocated to each of Consolidated Amalgamated's three businesses using each of the capital allocation methodologies discussed above. Depending on the methodology selected, the allocated equity capital, and thus the reported EVA^{SM}, of a business unit can vary dramatically.

Clearly the capital allocation methodology selected will affect not only the reported EVA^{SM} of each

Table 8

Equity Capital Allocations for Consolidated Amalgamated Bank, by Allocation Methodology

Business Unit	(1) Stand Alone: Peer Group ($millions)	(2) Stand Alone: Equal Probability of Insolvency ($millions)	(3) Scaled Diversification ($millions)	(4) Internal Betas ($millions)	(5) Marginal Capital ($millions)
Credit Card	2,018	2,018	822	1,526	767
Mortgage Banking	1,989	3,779	1,540	1,217	(683)
Subprime Lending	1,666	4,827	1,967	1,586	368
Unallocated Capital					3,877
Bank Total	5,633	10,624	4,329	4,329	4,329

Source: Column 1: Table 2
Column 2: Table 3
Column 3: Table 4
Column 4: Table 6
Column 5: Table 7

Internal Betas and Marginal Capital

Internal Betas

The risk of a bank (σ^2_{Bank}) with n different business units is given by the formula:

$$\sigma^2_{Bank} = \Sigma\Sigma w_i w_j cov_{i,j} \quad \text{(B-1)}$$

where w_i is the proportion of assets used by the i-th business unit, and $cov_{i,j}$ is the covariance of returns between the i-th and j-th business unit. This relationship is depicted in Table B-1 as the sum of the terms of a matrix of the business unit variances and covariances,[a] with each row representing a different business unit. Then the risk contribution of business 1 can be expressed as the sum of the terms in row 1, weighted by the assets of the business:

Risk contribution of business 1 =

$$w_1 \Sigma w_j cov_{1,j} = w_1 cov_{1, Bank} . \quad \text{(B-2)}$$

To measure the proportion of total risk contributed by business 1, we divide equation (B-2) by the overall risk of the bank:

Proportional risk contribution of business 1 =

$$w_1 cov_{1, Bank} / \sigma^2_{Bank} = w_1 \beta_1 . \quad \text{(B-3)}$$

But this is the internal beta of business 1. Because the proportion of risk accounted for by all the business units in the bank must equal the risk of the bank, then

$$\Sigma w_i \beta_i = 1 . \quad \text{(B-4)}$$

While the internal beta approach divides up the risk of the bank and does so in a way that incorporates the correlation in returns between the business unit and the bank, using the internal beta to allocate capital involves two very restrictive assumptions. First, because the risk of the bank is the weighted sum of the risk contribution of the business units, it already incorporates the risk contribution of business 1. That is, the risk contribution of each business is calculated on an ex post basis, assuming that the business is already and will remain a part of the bank. If a new business unit is added (deleted) then the variance/covariance matrix used to calculate the risk of the bank will have to add (delete) both a row and a column and the weights of the original

[a] Notice that the covariance of a variable with itself equals the variance of the variable.

Table B-1
Risk Contribution By Business Unit: The Internal Beta Approach

Business Unit	1	2	3		N	
1	$w_1^2\sigma_1^2$	$w_1w_2cov_{1,2}$	$w_1w_3cov_{1,3}$	—	$w_1w_ncov_{1,n}$	Risk Contribution = $w_1\Sigma w_j\, cov_{1,j} = w_1cov_{1,Bank}$
2	$w_2w_1cov_{1,2}$	$w_2^2\sigma_2^2$	$w_2w_3cov_{2,3}$	—	$w_2w_ncov_{2,n}$	Risk Contribution = $w_2\Sigma w_jcov_{2,j} = w_2cov_{2,Bank}$
3	$w_3w_1cov_{1,3}$	$w_3w_2cov_{2,3}$	$w_3^2\sigma_3^2$	—	$w_3w_ncov_{3,n}$	Risk Contribution = $w_3\Sigma w_j\, cov_{3,j} = w_3cov_{3,Bank}$
	—	—	—	—		—
N	$w_nw_1cov_{1,n}$	$w_nw_2cov_{2,n}$	$w_nw_3cov_{n,3}$	—	$w_n^2\sigma_n^2$	Risk Contribution = $w_n\Sigma w_jcov_{n,j} = w_ncov_{n,Bank}$
						Total Contribution = $\Sigma\Sigma\, w_iw_j\, cov_{i,j} = \sigma^2_{Bank}$

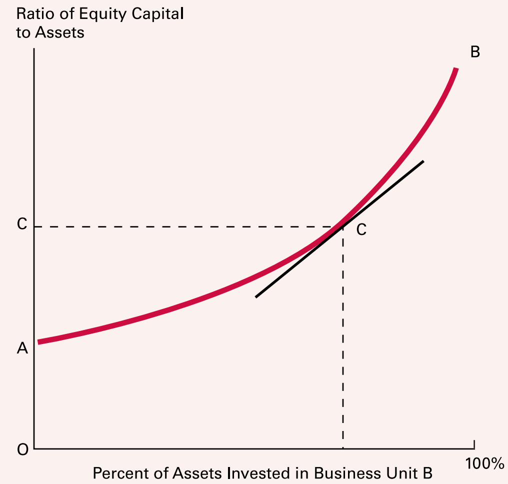
entries will change so that each row of the matrix, as well as the overall risk of the bank, will change. Second, the calculated risk contributions for each business unit are only valid for the asset weightings used. Any disproportional change in the relative importance of a business unit will change the weights on all of the entries in the variance/covariance matrix and thus result in a change not only in the internal betas of that business unit, but also in the internal betas of all of the other business units. Thus, capital allocations calculated using the internal beta approach are valid only for a specific mix of business units and cannot be used for other configurations of business units or asset weightings. Moreover, the capital allocation and reported EVA^{SM} of each business unit will be affected by the activity of the other business units in the bank.

Marginal Capital

Because a disproportionate change in the activity of one business unit affects the risk weighting of all of the business units, the incremental change in the total risk of the bank is not just the increment in the risk contribution of the particular business unit initiating the change, but also includes the effects on the risk contributions of all of the other business units in the variance/covariance matrix. Except in special circumstances this marginal risk contribution will not be equal to the risk contribution computed using internal betas. This is shown in Figure B-1 for a bank consisting of two business units. Business unit 1 is relatively low-risk and low-return, while business unit 2 is relatively high-risk, high-return. Figure B-1 shows the equity capital-to-asset ratio required to achieve a constant Z-ratio for different asset weightings of units 1 and 2. At point A, 100 percent of the bank's assets are comprised of unit 1 and the bank's required capital-to-asset ratio is simply the stand-alone required capital ratio for unit 1. At point B, 100 percent of the bank's assets are invested in unit 2, and the bank's required capital-to asset ratio is simply the stand-alone required capital ratio for unit 2. The curve AB represents the equity capital-to-asset ratios for all

Figure B-1

Diversification and Marginal Risk



the weightings of unit 1 and 2 to achieve the same probability of insolvency and is thus an iso-insolvency curve. It is convex because the returns of the businesses are assumed to be imperfectly positively correlated.

As shown in Figure B-1, each point on the iso-insolvency curve shows a different capital-to-asset ratio corresponding to a different mix of business units. If the bank increases the size of unit 2 relative to unit 1 it will move to the right along the curve and its required capital-to-asset ratio will increase. The rate at which the required capital-to-asset ratio increases is equivalent to the marginal capital ratio and can be shown as the slope of a tangent to the iso-insolvency curve. At point C, the required capital-to-asset ratio is OC, but the marginal capital is equal to the slope of the tangent at C, which is greater than OC. Thus, the marginal capital ratio will not equal the capital ratio for the bank as a whole, nor will it be a weighted average of the stand-alone risk of each of the business units.

business, but also how well the resulting measure captures the true economic contribution of the business and the incremental risk for the bank. For example, if a stand-alone methodology is selected, then the calculated EVA^{SM}s will be lower than if diversification effects are taken into account, and the sum of the business unit EVA^{SM}s will be less than the EVA^{SM} of the bank as a whole. If diversification effects are included by scaling down stand-alone allocations, then the unit EVA^{SM}s will sum to the bank's EVA^{SM}, but the EVA^{SM} of inefficient users of capital will be improved more than those of efficient users, and capital allocations will still fail to reflect the actual risk contributions of the businesses. In particular, the EVA^{SM}s of inefficient users of capital whose returns are highly correlated with the rest of the bank will be biased upward compared to the EVA^{SM}s of efficient users of capital with low positive or negative correlations.

If the bank chooses to allocate capital based on internal betas, then the business unit EVA^{SM}s will sum to that of the bank, and the capital allocation of each unit will reflect not only its stand-alone risk but also the interaction of the business with the other parts of the bank. But the capital ratios calculated using internal betas do not reflect the incremental risk associated with acquisitions, divestitures, or a change in the scale of operations, and thus will result in biased estimates of the associated incremental EVA^{SM}. Moreover, if the returns of the unit and the bank are negatively correlated, then the internal beta and capital allocation of the business would be negative.

While some observers have argued that negative capital allocations are nonsensical, in fact they merely reflect the reduction in required bank capital that occurs when a unit with negatively correlated returns is combined with the rest of the bank. Negative capital allocations can easily be incorporated into the EVA^{SM} equation shown at the beginning of Section II. The effect of such a negative equity capital allocation is to create a negative opportunity cost of capital and increase the EVA^{SM} of the unit so that it is greater than its adjusted earnings. This augmented EVA^{SM} reflects not only the earnings of the business but the saving in capital costs resulting from the unit's function as a natural hedge. However, a negative equity allocation to a particular business may represent a considerable challenge in terms of convincing the managers of the other business units that they have been treated fairly. Moreover, it is questionable whether one would wish to compensate the manager of a business with a negative equity allocation on the basis of the unit's EVA^{SM}, since the latter represents not only the economic profit of the business but also its value as a natural hedge, which has nothing to do with the manager's efforts.

Finally, if marginal capital is used as a basis for allocations, then the EVA^{SM}s of important strategic decisions will more accurately reflect their contribution to the bank. But as Merton and Perold (1995) have shown, the sum of the unit marginal capital allocations will be less than the capital of the bank, and the sum of the EVA^{SM}s of the businesses will be more than the EVA^{SM} of the bank. For example, as shown in Table 7, the allocations of marginal capital sum to only about 10 percent of the total capital of Amalgamated Consolidated, leaving about 90 percent of the bank's equity capital unallocated.[18] Thus it would be conceptually possible for each of the businesses to be generating a positive EVA^{SM}, but for the bank as a whole to be generating a zero or negative EVA^{SM}. Moreover, negative capital allocations are more likely to occur, resulting in the communications and compensation issues discussed above. Unfortunately, none of the capital allocation methodologies described above will result in an EVA^{SM} that will in all circumstances accurately reflect the economic contribution of the business unit.

V. Conclusion: Using Economic Profit to Measure Performance

Clearly, the incorporation of an opportunity cost of equity capital into a bank's performance measurement system potentially can offer great benefits in terms of improved risk management, greater efficiency in the use of capital, and quicker and more informed decision-making on the part of managers. But if business units are related, either operationally or in their use of equity capital, then the isolation of the earnings and economic capital used by each business becomes problematic. In such situations, estimates of economic profit may be biased and lead to poor decision-making. Attacking this problem is difficult because it is essentially a measurement issue: Identification of the extent and source of the problem would resolve it.

Are there rules of thumb that might help managers to assess the problem of relatedness in perfor-

[18] This unallocated capital is not excess, but represents the amount required to protect depositors and creditors against the positive correlation of returns among the individual business units.

mance measurement? If the assumption is made that the extent of relatedness can be approximated by the degree of correlation among the businesses in their returns, then we can distinguish between two situations: businesses with a high degree of relatedness and correlation in their earnings, and those with little or no relatedness or correlation.[19] In the case of the former, the earnings of the businesses are likely to be related. Thus, the economic contribution of units that generate positive externalities for other units is likely to be underestimated, leading to underinvestment in these units. On the other hand, the business units are unlikely to act as natural hedges for each other, so that each business will need approximately the equity capital required on a stand-alone basis. In this case managers need to focus on identifying cross-unit effects on revenue and expenses, but they can apply a relatively simple capital allocation scheme.

[19] Of course two businesses whose returns are highly correlated may be truly related, or they may be independent but their returns may be correlated with some third factor, such as interest rates or exchange rates. Thus, the effects of plausible third factors should be isolated before the conclusion is reached that the businesses are truly related.

In the case of units with low or negative correlations in returns, earnings will not be affected but the units will act as partial natural hedges, reducing the equity capital required for each unit. If the hedging effects of diversification are not taken into account, excessive equity capital will be allocated to these businesses, biasing downward reported economic profit and once again leading to underinvestment. In this case managers should focus on the capital allocation methodology, to ensure that the allocated capital is proportional to the actual risk contribution of the business.

Finally, while the concept of economic profit has powerful conceptual appeal, the ambiguities that surround its calculation indicate that no single measure of economic profit is able to capture all the subtle complexities, and that managers need to employ many specialized performance measures. For example, marginal capital might be used to compute the EVA^{SM} of a potential acquisition, but capital based on internal betas to measure the EVA^{SM} of an existing line of business. While the concept of economic profit may ultimately result in better measurement of bank performance, it is unlikely to simplify the measurement process.

References

Bacidore, Jeffrey M., John A. Boquist, Todd T. Milbourn, and Anjan V. Thakor. 1997. "EVA and Total Quality Management." *Journal of Applied Corporate Finance,* Summer, vol. 10, no. 2, pp. 81–89.

Brickley, James A., Clifford W. Smith, Jr., and Jerold L. Zimmerman. 1997. "Management Fads and Organizational Architecture." *Journal of Applied Corporate Finance,* Summer, vol. 10, no. 2, pp. 24–39.

Hannan, Timothy H. and Gerald A. Hanweck. 1988. "Bank Insolvency Risk and the Market for Large Certificates of Deposit." *Journal of Money, Credit and Banking*, May, pp. 203–11.

Kramer, Jonathan K. and George Pushner. 1997. "An Empirical Analysis of Economic Value Added as a Proxy for Market Value Added." *Financial Practice and Education*, Spring/Summer, vol. 7, no. 1, pp. 41–49.

Lehn, Kenneth and Anil Makhija. 1996. "EVA and MVA as Performance Measures and Signals for Strategic Change." *Strategy and Leadership*, May/June, pp. 34–38.

Merton, Robert C. and Andre F. Perold. 1995. "Theory of Risk Capital in Financial Firms." *The Journal of Applied Corporate Finance*, Fall, pp. 16–32.

Myers, Randy. 1996. "The Metric Wars." *CFO*, October, pp. 41–50.

O'Byrne, Stephen F. 1997. "EVA^{SM} and the Shareholder." *Financial Practice and Education*, Spring/Summer, vol. 7, no. 1, pp. 50–54.

Porter, Michael E. 1985. *Competitive Advantage.* New York: Free Press.

Prahalad, C.K. and Gary Hamel. 1990. "Core Competencies of the Corporation," *Harvard Business Review,* May, pp. 79–91.

Ross, Irwin. 1997. "The 1996 Stern Stewart Performance 1000." *Journal of Applied Corporate Finance,* Winter, vol. 9, no. 4, pp. 115–28.

Stern, Joel M., G. Bennet Stewart III, and Donald H. Chew, Jr. 1995. "The EVA Financial System." *Journal of Applied Corporate Finance,* Summer, pp. 38–55.

Stewart, G. Bennet III. 1991. *The Quest for Value: A Guide for Senior Managers.* New York: Harper Business.

Uyemura, Dennis G., Charles C. Kantor, and Justin M. Pettit. 1996. "EVA^{SM} For Banks: Value Creation, Risk Management, and Profitability Measurement." *Journal of Applied Corporate Finance,* Summer, vol. 9, no. 2. pp. 94–109.

Wruck, Karen Hopper and Michael C. Jensen. 1997. "Science, Specific Knowledge, and Total Quality Management," *Journal of Applied Corporate Finance,* Summer, vol. 10, no. 2, pp. 8–23.

Zimmerman, Jerold L. 1997. "EVA^{SM} and Divisional Performance Measurement: Capturing Synergies and Other Issues." *Journal of Applied Corporate Finance,* Summer, vol. 10, no. 2, pp. 98–109.